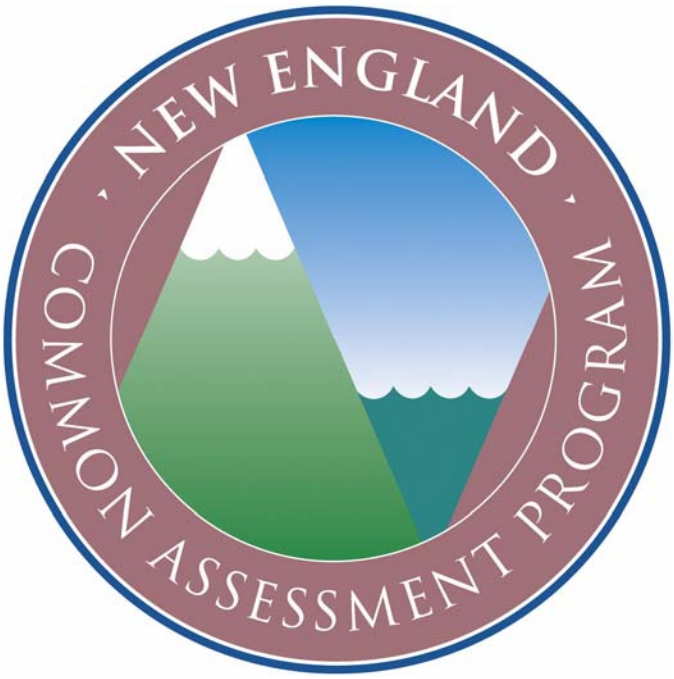

New England Common Test Program

2006–2007 Technical Report

August 2007

Measured Progress 100 Education Way, Dover, NH 03820 (800) 431-8901

TABLE OF CONTENTS

CHAPTER 1—OVERVIEW

1.1 PURPOSE OF THE NEW ENGLAND COMMON TEST PROGRAM

The New England Common Test Program (NECAP) is the result of collaboration among New Hampshire (NH), Rhode Island (RI), and Vermont (VT) to build a set of tests for grades 3 through 8 to meet the requirements of the No Child Left Behind Act (NCLB). The purposes of the tests are as follows: (1) Provide data on student achievement in reading/language arts and mathematics to meet the requirements of NCLB; (2) provide information to support program evaluation and improvement; and (3) provide to parents and the public information on the performance of students and schools. The tests are constructed to meet rigorous technical criteria, include universal design elements and accommodations so that students can access test content, and gather reliable student demographic information for accurate reporting. School improvement is supported by

- providing a transparent test design through the grade-level expectations (GLEs), distributions of emphasis, and practice tests
- reporting results by GLE subtopics, released items, and subgroups
- hosting test interpretation workshops to foster understanding of results

Student-level results are provided to schools and families to be used as one piece of evidence about progress and learning that occurred on the prior year's GLEs. The results are a status report of a student's performance against GLEs and should be used cautiously in partnership with local data.

1.2 PURPOSE OF THIS REPORT

The purpose of this report is to document the technical aspects of the 2006–2007 NECAP. In October of 2006, students in grades 3 through 8 participated in the administration of the NECAP in reading and mathematics. Students in grades 5 and 8 also participated in writing. This report provides information about the technical quality of those tests, including a description of the processes used to develop, administer, and score the tests and to analyze the test results. This report is intended to serve as

a guide for replicating and/or improving the procedures in subsequent years.

Though some parts of this technical report may be used by educated laypersons, the intended audience is experts in psychometrics and educational research. The report assumes a working knowledge of measurement concepts, such as "reliability" and "validity," and statistical concepts, such as "correlation" and "central tendency." In some chapters, the reader is presumed also to have basic familiarity with advanced topics in measurement and statistics.

1.3 ORGANIZATION OF THIS REPORT

The organization of this report is based on the conceptual flow of a test's life span; the report begins with the initial test specification and addresses all the intermediate steps that lead to final score reporting. Section I provides a description of the NECAP test. It consists of four chapters covering the test design and development process; the administration of the tests; scoring; and scaling and equating. Section II provides statistical and psychometric summaries. It consists of three chapters covering item analysis, reliability, and validity. Section III covers NECAP score reporting. Section IV contains references, and Section V contains the appendices.

SECTION I
DESCRIPTION OF THE 2006 NECAP TEST
CHAPTER 2—DEVELOPMENT AND TEST DESIGN

2.1 OPERATIONAL DEVELOPMENT PROCESS

GRADE-LEVEL EXPECTATIONS

NECAP test items are directly linked to *content standards* and *performance indicators* described in the GLEs. The content standards for each grade are grouped into content clusters for purposes of reporting results; the performance indicators are used by content specialists to help guide the development of test questions. An item may address one, several, or all of the performance indicators.

EXTERNAL ITEM REVIEW

Item Review Committees (IRCs) were formed by the states to provide an external review of items. The committees are made up of teachers, curriculum supervisors, and higher-education faculty from the states, and all committee members serve rotating terms. A list of IRC member names and affiliations is included in Appendix A. The committees review test items for the NECAP, provide feedback on the items, and make recommendations on which items should be selected for program use. The 2006–07 NECAP IRCs for each content area in grade levels 3 through 8 met in the spring of 2006. Committee members reviewed the entire set of embedded field-test items proposed for the 2007–08 operational test and made recommendations about selecting, revising, or eliminating specific items from the item pool for the operational test. Members reviewed each item against the following criteria:

- **Grade-Level Expectation Alignment**

  - Is the test item aligned to the appropriate GLE?

  - If not, which GLE or grade level is more appropriate?

- **Correctness**

  - Are the items and distracters correct with respect to content accuracy and developmental appropriateness?

  - Are the scoring guides consistent with GLE wording and developmental appropriateness?

- **Depth of Knowledge***

  - Are the items coded to the appropriate Depth of Knowledge?

  - If consensus cannot be reached, is there clarity around why the item might be on the borderline of two levels?

  * NECAP employed the work of Dr. Norman Webb to guide the development process with respect to Depth of Knowledge. Test specification documents identified ceilings and targets for Depth of Knowledge codings.

- **Language**

  - Is the item language clear?

  - Is the item language accurate (syntax, grammar, conventions)?

- **Universal Design**

  - Is there an appropriate use of simplified language (does not interfere with the construct being assessed)?

  - Are charts, tables, and diagrams easy to read and understandable?

  - Are charts, tables, and diagrams necessary to the item?

  - Are instructions easy to follow?

  - Is the item amenable to accommodations—read aloud, signed, or Braille?

INTERNAL ITEM REVIEW

- The lead Measured Progress test developer within the content specialty reviewed the formatted item, CR scoring guide, and any reading selections and graphics.
- The content reviewer considered item "integrity," item content and structure, appropriateness to designated content area, item format, clarity, possible ambiguity, answer cueing, appropriateness and quality of reading selections and graphics, and appropriateness of scoring guide descriptions and distinctions (as correlated to the item and within the guide itself). The item reviewer also ensured that, for each item, there was only one correct answer.
- The content reviewer also considered scorability and evaluated whether the scoring guide adequately addressed performance on the item.
- Fundamental questions that the content reviewer considered, but was not limited to, included the following:

  - What is the item asking?

  - Is the key the only possible key? (Is there only *one* correct answer?)

  - Is the CR item scorable as written (were the correct words used to elicit the response defined by the guide)?

  - Is the wording of the scoring guide appropriate and parallel to the item wording?

  - Is the item complete (e.g., with scoring guide, content codes, key, grade level, and identified contract)?

  - Is the item appropriate for the designated grade level?

BIAS AND SENSITIVITY REVIEW

Bias review is an essential component of the development process. During the bias review process, NECAP items were reviewed by a committee of teachers, English language learner (ELL) specialists, special-education teachers, and other educators and members of major constituency groups who represent the interests of legally protected and/or educationally disadvantaged groups. A list of bias

and sensitivity review committee member names and affiliations are included in Appendix A. Items were examined for issues that might offend or dismay students, teachers, or parents. Including such groups in the development of test items and materials can avoid many unduly controversial issues, and unfounded concerns can be allayed before the test forms are produced.

ITEM EDITING

Measured Progress editors reviewed and edited the items to ensure uniform style (based on *The Chicago Manual of Style,* 14th edition) and adherence to sound testing principles. These principles included the stipulation that items

- were correct with regard to grammar, punctuation, usage, and spelling
- were written in a clear, concise style
- contained unambiguous explanations to students as to what is required to attain a maximum score
- were written at a reading level that would allow the student to demonstrate his or her knowledge of the tested subject matter, regardless of reading ability
- exhibited high technical quality regarding psychometric characteristics
- had appropriate answer options or score-point descriptors
- were free of potentially sensitive content

REVIEWING AND REFINING

Test developers presented item sets to the development committees for their recommendations on which items should be available to include in the embedded field-test portions of the test. The NH, RI, and VT Departments of Education content specialists made the final selections with the assistance of Measured Progress at a final face-to-face meeting.

OPERATIONAL TEST ASSEMBLY

At Measured Progress, test assembly is the sorting and laying out of item sets into test forms. Criteria considered during this process included the following:

- **Content coverage/match to test design.** The Measured Progress curriculum and test specialists completed an initial sorting of items into sets based on a balance of content categories across sessions and forms, as well as a match to the test design (e.g., number of MC, SA, and CR items).
- **Item difficulty and complexity.** Item statistics drawn from the data analysis of previously tested items were used to ensure similar levels of difficulty and complexity across forms.
- **Visual balance.** Item sets were reviewed to ensure that each reflected a similar length and "density" of selected items (e.g., length/complexity of reading selections, number of graphics).
- **Option balance.** Each item set was checked to verify that it contained a roughly equivalent number of key options (A, B, C, and D).
- **Name balance.** Item sets were reviewed to ensure that a diversity of student names was used.
- **Bias.** Each item set was reviewed to ensure fairness and balance based on gender, ethnicity, religion, socioeconomic status, and other factors.
- **Page fit.** Item placement was modified to ensure the best fit and arrangement of items on any given page.
- **Facing-page issues.** For multiple items associated with a single stimulus (a graphic or reading selection), consideration was given both to whether those items needed to begin on a left- or right-hand page and to the nature and amount of material that needed to be placed on facing pages. These considerations served to minimize the amount of "page flipping" required of students.
- **Relationship between forms.** Although embedded field-test items differ from form to form, they must take up the same number of pages in each form so that sessions and content areas begin on the same page in every form. Therefore, the number of pages needed for the longest form often determines the layout of each form.

- **Visual appeal.** The visual accessibility of each page of the form was always taken into consideration, including such aspects as the amount of "white space," the density of the text, and the number of graphics.

EDITING DRAFTS OF OPERATIONAL TESTS

Any changes made by a test construction specialist must be reviewed and approved by a test developer. After a form had been laid out in what was considered its final form, it was reread to identify any final considerations, including the following:

- Editorial changes. All text was scrutinized for editorial accuracy, including consistency of instructional language, grammar, spelling, punctuation, and layout. Measured Progress's publishing standards are based on *The Chicago Manual of Style,* 14th edition.
- "Keying" items. Items were reviewed for any information that might "key" or provide information that would help to answer another item. Decisions about moving keying items are based on the severity of the "key-in" and the placement of the items in relation to each other within the form.
- Key patterns. The final sequence of keys was reviewed to ensure that their order appeared random (e.g., no recognizable pattern and no more than three of the same key in a row).

BRAILLE AND LARGE-PRINT TRANSLATION

Common items for grades 3 through 8 were translated into Braille by a subcontractor that specializes in test materials for blind and visually impaired students. In addition, Form 1 for each grade was also adapted into a large-print version.

2.2 ITEM TYPES

NH, RI, and VT educators and students were familiar with the item types that were used in the 2006–07 test, as all types had previously appeared on the 2005-06 NECAP. The item types used and the functions of each are described below.

Multiple-Choice (MC) items were administered in grades 3 through 8 in reading, mathematics,

and writing to provide breadth of coverage of the GLEs. Because they require approximately one minute for most students to answer, these items make efficient use of limited testing time and allow coverage of a wide range of knowledge and skills, including, for example, Word Identification (Word ID) and vocabulary skills.

Short-Answer (SA) items were administered in grades 3 through 8, mathematics only, to assess students' skills and their abilities to work with brief, well-structured problems that had one solution or a very limited number of solutions. SA items require approximately two to five minutes for most students to answer. The advantage of this item type is that it requires students to demonstrate knowledge and skills by generating, rather than merely selecting, an answer.

Constructed-Response (CR) items typically require students to use higher-order thinking skills—evaluation, analysis, summarization, and so on—in constructing a satisfactory response. CR items should take most students approximately five to ten minutes to complete. These items were administered in grades 3 through 8 in reading, in grades 5 and 8 in writing, and in grades 5 through 8 in mathematics.

A single common writing prompt with three SA planning box items was administered in grades 5 and 8. Students were given 45 minutes (plus limited additional time if necessary) to compose an extended response that was scored by two independent readers both on the quality of the stylistic and rhetorical aspects of the writing and on the use of standard English conventions. Students were encouraged to write a rough draft and were advised by the test administrator when to begin copying their final draft into their student answer booklets.

Approximately twenty-five percent of the common NECAP items were released to the public in 2006–07. The released NECAP items are posted on a Web site hosted by Measured Progress and on the Department of Education Web sites. Schools are encouraged to incorporate the use of released items in their instructional activities so that students will be familiar with them.

2.3 OPERATIONAL TEST DESIGNS AND BLUEPRINTS

Since the beginning of the program, the goal of the NECAP has been to measure what students know and are able to do by using a variety of test item types. The program was structured to use both common and matrix-sampled items. (Common items are those taken by all students at a given grade level; matrix-sampled items make up a pool that is divided among the multiple forms of the test at each grade level.) This design provides reliable and valid results at the student level and breadth of coverage of a content area for school results while minimizing testing time.) Note: Only common items are counted toward students' scaled scores.

EMBEDDED EQUATING ITEMS AND FIELD TEST

To ensure that NECAP scores obtained from different test forms and different years are equivalent to each other, a set of equating items is matrixed across forms of the reading and mathematics tests. (See Chapter 5 for more detail on the equating process.) Note: Equating items are not counted toward students' scaled scores.

The NECAP also includes embedded field test items in all content areas except writing. Because the field tested items are taken by many students, the sample is sufficient to produce reliable data with which to inform the process of selecting items for future tests. Embedding field tested items achieves two other objectives. First, it creates a pool of replacement items that, due to attrition caused by the release of common items each year, are needed in reading and mathematics. Second, embedding field-test items into the operational test ensures that students take the items under operational conditions. As with the matrixed equating items, field test items are not counted toward students' scaled scores.

TEST BOOKLET DESIGN

To accommodate the embedded equating and field test items in the 2006–07 NECAP, there were nine unique test forms at each grade. In all reading and mathematics test sessions, the equating and field-test items were distributed among the common items in a way that was not evident to test takers. The writing design called for one common test form that was made up of a single writing prompt with three

SA planning box items, four CR items, and ten MC items.

READING TEST DESIGN

Table 2-1 summarizes the numbers and types of items that were used in the NECAP reading test for 2006–07. Note that in reading, all students received the common items and one of either the equating or field test forms. Each MC item was worth one point, and each CR item was worth four points.

Table 2-1. 2006-07 NECAP Reading – Grades 3 through 8: Item Type and Numbers of Items.

Common – 2 long[1] and 2 short[1] passages plus 4 stand-alone MC[2]		Matrix – Equating Forms 1,2,3 1 long and 1 short passages plus 2 stand-alone MC		Matrix – FT[3] Forms 4-7 1 long and 1 short passages plus 2 stand-alone MC		Matrix – FT[3] Forms 8–9 3 short passages plus 2 stand-alone MC		Total per student – 3 long and 3 short or 2 long and 5 short passages plus 6 stand-alone MC	
MC[2]	CR[2]	MC	CR	MC	CR	MC	CR	MC	CR
28	6	14	3	14	3	14	3	42	9

[1]Long passages have 8 MC and 2 CR items; short passages have 4 MC and 1 CR items
[2]MC = multiple choice; CR = constructed response
[3]FT = field test

READING BLUEPRINT

As indicated earlier, the test framework for reading was based on the *NECAP Grade Level Expectations*, and all items on the NECAP test were designed to measure a specific GLE. The reading passages on the NECAP test are broken down into the following categories:

- **Literary passages** presenting a variety of forms: modern narratives; diary entries; drama; poetry; biographies; essays; excerpts from novels; short stories; and traditional narratives, such as fables, tall tales, myths, and folktales.
- **Informational passages** that are factual texts and often deal with the areas of science and social studies. These passages are taken from such sources as newspapers, magazines, and book excerpts. Informational text could also be directions, manuals, and recipes, etc.

  The passages are authentic texts—selected from grade-level-appropriate reading sources—that students would be likely to experience in both the classroom and independent reading. Passages are written specifically for the test; all are collected from published works.

Reading comprehension is assessed by items on the NECAP test that are *dually*-categorized by

the type of passage associated and the level of comprehension measured. The level of comprehension is designated as either "Initial Understanding" or "Analysis and Interpretation." Word identification and vocabulary skills are assessed at each grade level primarily through MC items. The distribution of emphasis for reading is shown in Table 2-2.

Table 2-2. 2006-07 NECAP Reading – Grades 3 through 8: Distribution of Emphasis by Grade (in targeted percentage of test).

Emphasis	GLE grade (grade tested)					
	2 (3)	3 (4)	4 (5)	5 (6)	6 (7)	7 (8)
Word Identification Skills and Strategies	20%	15%	0%	0%	0%	0%
Vocabulary Strategies/Breadth of Vocabulary	20%	20%	20%	20%	20%	20%
Initial Understanding of Literary Text	20%	20%	20%	20%	15%	15%
Initial Understanding of Informational Text	20%	20%	20%	20%	20%	20%
Analysis and Interpretation of Literary Text	10%	15%	20%	20%	25%	25%
Analysis and Interpretation of Informational Text	10%	10%	20%	20%	20%	20%
Total	100%	100%	100%	100%	100%	100%

Table 2-3 shows the subcategory reporting structure for reading and the maximum possible number of raw score points that students could earn. (With the exception of Word ID/Vocabulary items, reading items were reported in two ways: type of text and level of comprehension.)

Table 2-3. 2006-07 NECAP Reading – Grades 3 through 8: Reporting Subcategories and Possible Raw Score Points by Grade.

Subcategory		Grade Tested					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Word ID/ Vocabulary		20	20	10	10	10	10
Type of Text	Literary	16	16	21	21	21	22
	Informational	16	16	21	21	21	20
Level of Comprehension	Initial Understanding	20	20	26	21	17	18
	Analysis and Interpretation	12	12	16	21	25	24
	Total	52[1]	52	52	52	52	52

[1]Total possible points in reading is the points in Word ID/Vocabulary plus either Type of Text or Level of Comprehension (comprehension items are dually-categorized by type of text and level of comprehension).

Table 2-4 lists the percentage of total score points assigned to each level of Depth of Knowledge in Reading.

Table 2-4. 2006-07 NECAP Reading – Grades 3 through 8: Depth of Knowledge (DOK) by Grade (in percentage of test).

	Grade Tested					
DOK	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**
Level 1	34%	27%	15%	17%	15%	17%
Level 2	58%	65%	70%	58%	44%	52%
Level 3	8%	8%	15%	25%	41%	31%
Total	100%	100%	100%	100%	100%	100%

MATHEMATICS TEST DESIGN

Table 2-5 summarizes the numbers and types of items that were used in the NECAP mathematics test in grades 3 and 4 for 2006–07. Table 2-6 summarizes the numbers and types of items that were used in the NECAP mathematics test in grades 5 through 8 for 2006–07. Note that all students received the common, in one of the forms that also includes equating and field test items. Each MC item was worth one point, each SA item either one or two points, and each CR item four points. Score points within a grade level were evenly divided, so that MC items represented approximately fifty percent of possible score points, and SA and CR items together represented approximately fifty percent of score points.

Table 2-5. 2006-07 NECAP Mathematics – Grades 3 and 4: Item Type and Numbers of Items.

Common			Matrix – Equating			Matrix – FT[2]			Total per Student		
MC[1]	SA1[1]	SA2[1]	MC	SA1	SA2	MC	SA1	SA2	MC	SA1	SA2
35	10	10	6	2	2	3	1	1	44	13	13

[1]MC = multiple choice; SA1 = 1-point short answer; SA2 = 2-point short answer
[2]FT = field test

Table 2-6. 2006-07 NECAP Mathematics – Grades 5 through 8: Item Type and Numbers of Items.

Common				Matrix – Equating				Matrix – FT[2]				Total per Student			
MC[1]	SA1[1]	SA2[1]	CR[1]	MC	SA1	SA2	CR	MC	SA1	SA2	CR	MC	SA1	SA2	CR
32	6	6	4	6	2	2	1	3	1	1	1	41	9	9	6

[1]MC = multiple choice; SA1 = 1-point short answer; SA2 = 2-point short answer; CR = constructed response
[2]FT = field test

THE USE OF CALCULATORS ON THE NECAP

The mathematics specialists from the NH, RI, and VT Departments of Education who designed the mathematics test acknowledge the importance of mastering arithmetic algorithms. At the same time, they understand that the use of calculators is a necessary and important skill. Calculators can save time and prevent error in the measurement of some higher-order thinking skills, allowing students to work more sophisticated and intricate problems. For these reasons, it was decided that calculators should be permitted in the first of the three sessions of the NECAP mathematics test and prohibited in the remaining two sessions (test sessions are discussed at the end of this chapter).

MATHEMATICS BLUEPRINT

The test framework for mathematics was based on the *NECAP Grade Level Expectations*, and all items on the NECAP test were designed to measure a specific GLE. The mathematics items are organized into four content standards as shown on the following list:

- **Numbers and Operations:** Students understand and demonstrate a sense of what numbers mean and how they are used. Students understand and demonstrate computation skills.
- **Geometry and Measurement:** Students understand and apply concepts from geometry. Students understand and demonstrate measurement skills.

- **Functions and Algebra:** Students understand that mathematics is the science of patterns, relationships, and functions. Students understand and apply algebraic concepts.
- **Data, Statistics, and Probability:** Students understand and apply concepts of data analysis. Students understand and apply concepts of probability.

In addition, problem solving, reasoning, connections, and communication are embedded throughout the GLEs. The distribution of emphasis for Mathematics is shown in Table 2-7.

Table 2-7. 2006-07 NECAP Mathematics – Grades 3 through 8: Distribution of Emphasis (in targeted percentage of test).

Emphasis	GLE grade (grade tested) 2 (3)	3 (4)	4 (5)	5 (6)	6 (7)	7 (8)
Numbers and Operations	55%	50%	45%	40%	30%	20%
Geometry and Measurement	15%	20%	20%	25%	25%	25%
Functions and Algebra	15%	15%	20%	20%	30%	40%
Data, Statistics, and Probability	15%	15%	15%	15%	15%	15%
Total	100%	100%	100%	100%	100%	100%

Table 2-8 shows the subcategory reporting structure for writing and the maximum possible number of raw score points that students could earn. It can be seen that the goal for distribution of score points, or balance of representation across the four content strands, varies from grade to grade. Note: Only common items are counted toward students' scaled scores.

Table 2-8. 2006-07 NECAP Mathematics – Grades 3 through 8: Reporting Subcategories and Possible Raw Score Points by Grade.

Subcategory	Grade Tested Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Numbers and Operations	35	32	30	26	20	13
Geometry and Measurement	10	13	13	17	16	16
Functions and Algebra	10	10	13	13	19	27
Data, Statistics, and Probability	10	10	10	10	11	10
Total	65	65	66	66	66	66

Table 2-9 lists the percentage of total score points assigned to each level of Depth of Knowledge in mathematics.

Table 2-9. 2006-07 NECAP Mathematics – Grades 3 through 8: Depth of Knowledge (DOK) by Grade (in percentage of test).

DOK	Grade Tested					
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Level 1	29%	24%	20%	17%	24%	20%
Level 2	63%	62%	63%	70%	59%	62%
Level 3	8%	14%	17%	13%	17%	18%
Total	100%	100%	100%	100%	100%	100%

WRITING TEST DESIGN

Table 2-10 summarizes the numbers and types of items that were used in the NECAP writing test in grades 5 and 8 for 2006–07. Note that all items on the writing test were common. Each MC item was worth one point, each CR item four points, each SA item one point, and the writing prompt 12 points.

Table 2-10. 2006-07 NECAP Writing – Grades 5 and 8: Item Type and Numbers of Items.

All Common – Total Per Student			
MC[1]	CR[1]	SA1[1]	WP[1]
10	3	3	1

[1]MC = multiple choice; CR = constructed response; SA1 = 1-point short answer; WP = Writing Prompt

WRITING BLUEPRINT

The test framework for writing was based on the *NECAP Grade Level Expectations*, and all items on the NECAP test were designed to measure a specific GLE. The content standards in writing identify four major genres that are assessed in the writing portion of the NECAP test each year.

- Writing in response to literary text
- Writing in response to informational text
- Narratives
- Informational writing (report/procedure for Grade 5 and persuasive at Grade 8)

The writing prompt and the three CR items each address a different genre. In addition, structures and conventions of language are assessed through MC items and throughout the student's writing. The prompts and CR items were developed with the following criteria as guidelines:

- the prompts must be interesting to students
- the prompts must be accessible to all students (i.e., all students would have something to say about the topics)
- the prompts must generate sufficient text to be effectively scored

The subcategory reporting structure for writing is shown in Table 2-11. Also displayed are the maximum possible number of raw score points that students could earn. The subcategory "Short Responses" lists the total raw score points from the three CR items; the subcategory "Extended Response" lists the total raw score points from the three SA items and the writing prompt.

Table 2-11. 2006-07 NECAP Writing– Grades 5 and 8: Reporting Subcategories and Possible Raw Score Points by Grade.

Subcategory	Grade Tested	
	Grade 5	Grade 8
Structures of Language and Writing Conventions	10	10
Short Responses	12	12
Extended Response	15	15
Total	37	37

Table 2-12 lists the percentage of total score points assigned to each level of Depth of Knowledge in writing.

Table 2-12. 2006-07 NECAP Writing – Grades 5 and 8: Depth of Knowledge (DOK) by Grade (in percentage of test).

DOK	Grade Tested	
	Grade 5	Grade 8
Level 1	19%	22%
Level 2	41%	38%
Level 3	40%	40%
Total	100%	100%

TEST SESSIONS

The NECAP tests were administered to grades 3 through 8 during October 2–24 in 2006. Schools were able to schedule testing sessions at any time during two weeks of this period, provided they followed the sequence in the scheduling guidelines detailed in test administration manuals and that all testing classes within a school were on the same schedule. A third week was reserved for make-up testing of students who were absent from initial test sessions.

The timing and scheduling guidelines for the NECAP tests were based on estimates of the time it would take an average student to respond to each type of item that makes up the test:

- multiple-choice – 1 minute
- short-answer (1 point) – 1 minute
- short-answer (2 point) – 2 minutes
- constructed-response – 10 minutes
- long writing prompt – 45 minutes

For the reading tests, the scheduling guidelines included an estimate of 10 minutes to read the stimulus material used in the test. Tables 2-13 through 2-16 show the distribution of items across the test sessions for each content area and grade levels.

Table 2-13. 2006-07 NECAP Reading – Grades 3 through 8: Test Sessions by Item Type.

Item Type[1]	Session 1 1 long and 1 short passage plus 2 stand-alone MC	Session 2 1 long and 1 short passage plus 2 stand-alone MC	Session 3 1 long and 1 short passage plus 2 stand-alone MC
MC	14	14	14
CR	3	3	3

[1]MC = multiple choice; CR = constructed response

Table 2-14. 2006-07 NECAP Mathematics – Grades 3 and 4: Test Sessions by Item Type.

Item Type[1]	Session 1	Session 2	Session 3
MC	15	15	14
SA1	4	3	6
SA2	4	5	4

[1]MC = multiple choice; SA1 = 1-point short answer; SA2 = 2-point short answer

Table 2-15. 2006-07 NECAP Mathematics – Grades 5 through 8: Test Sessions by Item Type.

Item Type[1]	Session 1	Session 2	Session 3
MC	14	14	13
SA1	3	3	3
SA2	3	3	3
CR	2	2	2

[1]MC = multiple choice; SA1 = 1-point short answer; SA2 = 2-point short answer; CR = constructed response

Table 2-16. 2006-07 NECAP Writing – Grades 5 and 8: Test Sessions by Item Type.

Item Type[1]	Session 1	Session 2
MC	10	0
CR	3	0
SA	0	3
WP	0	1

[1]MC = multiple choice; CR = constructed response; SA1 = 1-point short answer; WP = Writing Prompt

Though the guidelines for scheduling are based on the assumption that most students will complete the test within the time estimated, each test session was scheduled so that additional time was provided for students who needed it. Up to one-hundred percent additional time was allocated for each session (e.g., a 50-minute session could be extended by an additional 50 minutes).

If classroom space was not available for students who required additional time to complete the tests, schools were allowed to consider using another space for this purpose, such as the guidance office. If additional areas were not available, it was recommended that each classroom used for test administration be scheduled for the maximum amount of time. Detailed instructions on test administration and scheduling were provided in the test coordinators' and administrators' manuals.

Chapter 3—Test Administration

3.1 Responsibility for Administration

As indicated in the *Principal/Test Coordinator Manual*, principals and/or their designated NECAP test coordinator were responsible for the proper administration of the NECAP. Manuals that contained explicit directions and scripts to be read aloud to students by test administrators were used in order to ensure the uniformity of administration procedures from school to school.

3.2 Administration Procedures

Principals and/or their school's designated NECAP coordinator were instructed to read the *Principal/Test Coordinator Manual* before testing and to be familiar with the instructions provided in the *Test Administrator Manual*. The *Principal/Test Coordinator Manual* provided each school with checklists to help them to prepare for testing. The checklists outlined tasks to be performed by school staff before, during, and after test administration. Besides these checklists, the *Principal/Test Coordinator Manual* described the testing material being sent to each school and how to inventory the material, track it during administration, and return it after testing was complete. The *Test Administrator Manual* included checklists for the administrators to prepare themselves, their classrooms, and the students for the administration of the test. The *Test Administrator Manual* contained sections that detailed the procedures to be followed for each test session, and instructions for preparing the material before the principal/test coordinator would return it to Measured Progress.

3.3 PARTICIPATION REQUIREMENTS AND DOCUMENTATION

The legislation's intent is for *all* students in grades 3 though 8 to participate in the NECAP through standard administration, administration with accommodations, or alternate test. Furthermore, any student who is absent during any session of the NECAP is expected to take a makeup test within the three-week testing window.

Schools were required to return a student answer booklet for every enrolled student in the grade level. On those occasions when it was deemed impossible to test a particular student, school personnel were required to inform their Department of Education. The states included a grid on the student answer booklets that listed the approved reasons why a student answer booklet could be returned blank for one or more sessions of the test:

- **Student completed the Alternate Test for the 2005–2006 school year**
  If a student completed the alternate test in the previous school year, the student was not required to participate in the NECAP in 2006-07.
- **Student is new to the United States after October 1, 2005 and is LEP (reading and writing only)**
  First-year LEP students that took the ACCESS test of English language proficiency, as scheduled in their states, were not required to take the reading and writing tests in 2006–07. However, these students were required to take the mathematics test in 2006–07.
- **Student withdrew from school after October 1, 2006**
  If a student withdrew after October 1, 2006 but before completing all of the test sessions, school personnel were instructed to code this reason on the student's answer booklet.
- **Student enrolled in school after October 1, 2006**
  If a student enrolled after October 1, 2006 and was unable to complete all of the test sessions before the end of the testing administration window, school personnel were instructed to code this reason on the student's answer booklet.

- **State-approved special consideration**

  Each state department of education had a process for documenting and approving circumstances that made it impossible or not advisable for a student to participate in testing. Schools were required to obtain state approval before beginning testing.

- **Student was enrolled in school on October 1, 2006 and did not complete test for reasons other than those listed above**

  If a student was not tested for a reason not stated above, school personnel were instructed to code this reason on the student's answer booklet. These "Other" categories were considered "not state-approved."

Tables 3-1, 3-2, and 3-3 list the participation rates of the three states combined in reading, mathematics, and writing.

Table 3-1. Participation Rates for 2006-07 NECAP Reading.

Category	Description	Enrollment	Not Tested State-Approved	Not Tested Other	Number Tested	Percent Tested
All	All Students	205675	2903	1554	201218	98
Gender	Male	106334	1725	927	103682	98
	Female	99205	1157	616	97432	98
	Not Reported	136	21	11	104	76
Ethnicity	Am. Indian	997	17	14	966	97
	Asian	4713	117	50	4546	96
	Black	8623	220	119	8284	96
	Hispanic	16130	604	253	15273	95
	NHPI	558	10	7	541	97
	White	173694	1895	1082	170717	98
	Not Reported	960	40	29	891	93
LEP	Current	6055	678	170	5207	86
	Monitoring Year 1	1042	4	6	1032	99
	Monitoring Year 2	848	2	5	841	99
	Other	197730	2219	1373	194138	98
IEP	IEP	33596	1766	644	31186	93
	Other	172079	1137	910	170032	99
SES	SES	57894	1338	585	55971	97
	Other	147781	1565	969	145247	98
Migrant	Migrant	162	7	1	154	95
	Other	205513	2896	1553	201064	98
Title 1	Title 1	32323	623	317	31383	97
	Other	173352	2280	1237	169835	98
Plan 504	Plan 504	963	4	3	956	99
	Other	204712	2899	1551	200262	98

Table 3-2. Participation Rates for 2006-07 NECAP Mathematics.

Category	Description	Enrollment	Not Tested State-Approved	Not Tested Other	Number Tested	Percent Tested
All	All Students	205675	2306	1548	201821	98
Gender	Male	106334	1400	921	104013	98
	Female	99205	887	617	97701	98
	Not Reported	136	19	10	107	79
Ethnicity	Am. Indian	997	16	15	966	97
	Asian	4713	44	41	4628	98
	Black	8623	126	122	8375	97
	Hispanic	16130	227	197	15706	97
	NHPI	558	10	5	543	97
	White	173694	1846	1140	170708	98
	Not Reported	960	37	28	895	93
LEP	Current	6055	77	81	5897	97
	Monitoring Year 1	1042	3	6	1033	99
	Monitoring Year 2	848	2	5	841	99
	Other	197730	2224	1456	194050	98
IEP	IEP	33596	1770	668	31158	93
	Other	172079	536	880	170663	99
SES	SES	57894	915	555	56424	97
	Other	147781	1391	993	145397	98
Migrant	Migrant	162	6	1	155	96
	Other	205513	2300	1547	201666	98
Title 1	Title 1	29570	335	278	28957	98
	Other	176105	1971	1270	172864	98
Plan 504	Plan 504	963	4	4	955	99
	Other	204712	2302	1544	200866	98

Table 3-3. Participation Rates for 2006-07 NECAP Writing.

Category	Description	Enrollment	Not Tested State-Approved	Not Tested Other	Number Tested	Percent Tested
All	All Students	69562	995	774	67793	97
Gender	Male	36000	576	475	34949	97
	Female	33519	412	295	32812	98
	Not Reported	43	7	4	32	74
Ethnicity	Am. Indian	306	3	5	298	97
	Asian	1517	33	16	1468	97
	Black	2913	73	61	2779	95
	Hispanic	5292	195	129	4968	94
	NHPI	259	4	6	249	96
	White	58985	672	544	57769	98
	Not Reported	290	15	13	262	90
LEP	Current	1800	209	65	1526	85
	Monitoring Year 1	311	1	2	308	99
	Monitoring Year 2	265	1	2	262	99
	Other	67186	784	705	65697	98
IEP	IEP	11760	585	375	10800	92
	Other	57802	410	399	56993	99
SES	SES	18843	436	309	18098	96
	Other	50719	559	465	49695	98
Migrant	Migrant	53	2	0	51	96
	Other	69509	993	774	67742	97
Title 1	Title 1	9798	202	152	9444	96
	Other	59764	793	622	58349	98
Plan 504	Plan 504	363	2	2	359	99
	Other	69199	993	772	67434	97

3.4 ADMINISTRATOR TRAINING

In addition to distributing the *Principal/Test Coordinator* and *Test Administrator Manual*s, the NH, RI, and VT Departments of Education, along with Measured Progress, conducted test administration workshops in five separate regional locations in each state to inform school personnel about the NECAP and to provide training on the policies and procedures regarding administration of the NECAP tests.

3.5 DOCUMENTATION OF ACCOMMODATIONS

The *Principal/Test Coordinator* and *Test Administrator Manual* provided directions for coding the information related to accommodations and modifications on page 2 of the student answer booklet. All accommodations used during any test session were required to be coded in by authorized school personnel—not students—after testing was completed.

An *Accommodations, Guidelines, and Procedures: Administrator Training Guide* was also produced to provide detailed information on planning and implementing accommodations. This guide can be located on each state's Department of Education Web site. The states collectively made the decision that accommodations be made available to all students based on individual need regardless of disability status. Decisions regarding accommodations were to be made by the students' educational team on an individual basis and were to be consistent with those used during the students' regular classroom instruction. Making accommodations decisions on an entire-group basis rather than on an individual basis was not permitted. If the decision made by a student's educational team required an accommodation not listed in the state-approved Table of Standard Test Accommodations, schools were instructed to contact the Department of Education in advance of testing for specific instructions for coding the "Other Accommodations (E)" and/or "Modifications (F)" section.

Tables 3-4 and 3-5 show the accommodations observed for the October 2006 NECAP administration. The accommodation codes are defined in the Table of Standard Test Accommodations, which can be found in Appendix B. Information on the appropriateness and impact of accommodations may be found in Appendix M.

Table 3-4. 2006-07 NECAP – Grades 3 through 5: Accommodation Frequencies by Subject Area.

	Grade 3		Grade 4		Grade 5		
Accommodation.	Reading	Math	Reading	Math	Reading	Math	Writing
A01	634	643	682	672	627	613	575
A02	3671	3784	3943	4080	4039	4187	4015
A03	1229	1262	1364	1343	1138	1135	1088
A04	331	298	268	264	278	275	269
A05	8	8	6	5	13	9	9
A06	9	8	14	14	37	35	35
A07	1532	1527	1744	1722	1782	1816	1759
A08	1140	1185	1233	1290	1080	1113	1053
A09	3	1	2	2	7	8	6
B01	231	224	186	182	224	224	210
B02	1901	1841	1971	1937	2098	2096	1973
B03	2133	2189	2342	2470	2999	3061	2562
C01	3	3	3	3	3	3	3
C02	46	43	31	32	38	38	35
C03	15	15	10	12	31	33	30
C04	0	3280	0	3373	0	2927	2680
C05	554	430	504	406	428	341	335
C06	20	47	35	72	44	53	25
C07	590	523	670	612	644	584	524
C08	6	6	4	6	9	11	8
C09	232	159	227	162	168	115	116
C10	22	7	12	6	19	11	10
C11	58	49	51	41	34	35	33
C12	0	14	0	29	0	24	12
C13	0	4	0	1	0	3	0
D01	22	18	26	20	96	54	159
D02	64	63	86	66	121	90	126
D03	3	4	5	8	10	6	6
D04	98	103	79	84	137	128	103
D05	828	781	871	823	745	610	0
D06	6	7	15	13	22	18	0
E01	4	6	2	5	4	6	4
E02	0	0	0	0	0	0	34
F01	0	19	0	70	0	25	0
F02	23	0	16	0	14	0	0
F03	3	4	2	3	3	5	10

Table 3-5. 2006-07 NECAP – Grades 6 through 8: Accommodation Frequencies by Subject Area.

	Grade 6		Grade 7		Grade 8		
Accommodation.	Reading	Math	Reading	Math	Reading	Math	Writing
A01	472	466	388	335	348	340	321
A02	3905	3957	3779	3818	3529	3549	3448
A03	945	933	632	609	497	489	473
A04	311	305	209	195	196	200	185
A05	8	5	11	4	14	19	12
A06	13	11	15	19	13	15	12
A07	1749	1717	1640	1592	1514	1517	1487
A08	672	690	507	526	383	394	361
A09	10	13	7	5	7	9	7
B01	193	191	156	150	176	172	171
B02	1796	1776	1338	1313	1155	1149	1099
B03	2442	2483	2129	2223	1790	1860	1622
C01	0	0	0	0	2	2	2
C02	23	23	20	17	21	25	20
C03	5	6	11	14	26	25	22
C04	0	2066	0	1487	0	1413	1333
C05	203	187	101	74	98	87	82
C06	59	91	24	42	37	65	39
C07	523	517	288	285	254	242	230
C08	12	12	13	10	4	6	4
C09	84	65	31	28	8	7	10
C10	2	1	4	3	5	6	7
C11	26	26	8	8	5	5	5
C12	0	80	0	70	0	65	51
C13	0	0	0	1	0	1	0
D01	127	62	141	63	169	89	229
D02	59	44	60	47	48	35	55
D03	9	9	8	8	3	4	1
D04	98	91	77	80	62	60	46
D05	540	392	335	252	251	214	0
D06	20	13	15	9	8	8	0
E01	0	0	0	1	5	5	5
E02	0	0	0	0	0	0	24
F01	0	99	0	63	0	55	0
F02	19	0	17	0	21	0	0
F03	26	26	2	2	5	4	5

3.6 TEST SECURITY

Maintaining test security is critical to the success of the New England Common Test program and the continued partnership among the three states. The *Principal/Test Coordinator Manual* and the *Test Administrator Manual*s explain in detail all test security measures and test administration procedures. School personnel were informed that any concerns about breaches in test security were to be reported to the schools' test coordinator and principal immediately. The test coordinator and/or principal were responsible for immediately reporting the concern to the district superintendent and the state director of testing at the department of education. Test Security was also strongly emphasized at test administration workshops that were conducted in all three states.

The three states also required the principal of each school that participated in testing to log on to a secure website to complete the *Principal's Certification of Proper Test Administration* form for each grade level tested. Principal's were requested to provide the number of secure tests received from Measured Progress, the number of tests administered to students, and the number of secure test materials that they were returning to Measured Progress. Principals were then instructed to print off a hard copy of the form, sign it, and return it with their test materials shipment. By signing the form, the principal was certifying that the tests were administered according to the test administration procedures outlined in the *Principal/Test Coordinator* and *Test Administrator* Manuals, that they maintained the security of the tests, that no secure material was duplicated or in any way retained in the school, and that all test materials had been accounted for and returned to Measured Progress.

3.7 TEST ADMINISTRATION WINDOW

The test administration window was October 2–24, 2006.

3.8 NECAP SERVICE CENTER

To provide additional support to schools before, during, and after testing, Measured Progress established the NECAP Service Center. The additional support that the Service Center provides is an essential element to the successful administration of any statewide test program. It provides a centralized location to which individuals in the field can call using a toll-free number and ask specific questions or report any problems they may be experiencing.

The Service Center was staffed by representatives at varying levels based on need volume and was available from 8:00 AM to 4:00 PM beginning two weeks before the start of testing and ending two weeks after testing. The representatives were responsible for receiving, responding to, and tracking calls, then routing issues to the appropriate person(s) for resolution. All calls were logged into a database that was provided to each state after testing was completed.

CHAPTER 4—SCORING

4.1 IMAGING PROCESS

When the 2006–07 NECAP student answer booklets arrived at Measured Progress, they were logged in, identified with pre-printed scannable school information header sheets, examined for extraneous materials, and batched. They were then moved to the scanning area for imaging. Booklets were scanned and all necessary information to produce required reports was captured and converted into an electronic format (e.g., all student identification and demographics, CR answers, and digital image clips of hand-written writing-prompt responses). Such digital image-clip information allows Measured Progress to replicate student responses, just as they appeared originally, onto readers' monitors for scoring. All remaining processes—data processing, benchmarking, scoring, data analysis, and reporting—are accomplished without further reference to original paper forms.

The first step in digitally converting student booklets was removal of booklet bindings so that individual pages could pass through the scanners one at a time. Once booklets were cut, their pages were put back into their proper boxes and placed in storage until needed for scanning and imaging.

Customized scanning programs were prepared to selectively read the 2006-07 NECAP student answer booklets and to format the scanned information electronically according to pre-determined requirements. All information (including MC response data) that had been designated time-critical or process-critical was handled first.

4.2 QUALITY CONTROL

The scanning system used at Measured Progress is equipped with many built-in safeguards that prevent data errors (e.g., real-time quality control checks, duplex reading). Furthermore, scanner hardware is continually monitored automatically, and if standards are not met, an error message is displayed and scanning shuts down. Areas automatically monitored include document page and integrity checks as well as internal checks of electronic functioning.

Before each scanning shift began, Measured Progress operators performed a diagnostic routine. In the event any inconsistencies were identified, an operator calibrated the machine and performed the test again. If the machine was still not up to standard, a field service engineer was called for assistance.

As a final safeguard, bubble-by-bubble and image-by–image spot checks of scanned files were routinely made throughout scanning runs to ensure data integrity.

After data were entered and scanning logs and paperwork completed, student booklets were put into storage (where they are kept for a minimum of 180 days beyond the close of the fiscal year). Once it had been determined that the 2006-07 NECAP databases were complete and accurate, batches were uploaded to Measured Progress' local area network (LAN).These data were then available to be scored or transferred as appropriate to the Internet, CD-ROM, or optical disk.

4.3 Hand-Scoring

iSCORE

Student responses to open-ended items on the 2006-07 NECAP were accessed as stored images off the LAN by qualified readers at computer terminals for "hand-scoring." All scoring personnel are subject to the same nondisclosure requirements and supervision as is regular Measured Progress staff.

Readers evaluate each response and record each student's score via keypad or mouse entry through the Measured Progress proprietary *iScore* system. All *iScore* scoring is "anonymous." No student names or scores are associated with viewed responses. Readers can only access student responses for items they are qualified to score. When a scorer finishes evaluating a response, another

random response immediately appears onscreen. In these ways, complete anonymity and randomization of student responses is ensured.

SCORER QUALIFICATIONS

Under the Director of Scoring Services, scoring staff carried out the various scoring operations. Scoring staff included

- chief readers (CRs), who oversaw all training and scoring within particular content areas;
- quality assurance coordinators (QACs), who led range finding and training activities and monitored scoring consistency and rates;
- senior readers (SRs), who performed read-behinds of readers and assisted at scoring tables as necessary; and
- readers, who performed the bulk of the scoring.

Table 4-1 summarizes the qualifications of the 2005-06 NECAP quality assurance coordinators and readers.

Table 4-1. 2006-07 NECAP QAC[1] and Reader Qualifications.

Scoring Responsibility	Educational Credentials				Total
	Doctorate	Masters	Bachelors	Other	
QAC	2.5%	35.0%	60.0%	2.5%	100%
Reader	4.0%	28.0%	61.0%	7.0%	100%

[1]QAC = Quality Assurance Coordinator

BENCHMARKING

Before the scheduled start of scoring activities, Measured Progress scoring center staff and test developers reviewed test items and scoring guides for benchmarking. One or two anchor examplars were selected for each item score point to prepare an anchor pack; an additional six to ten responses were selected to go into the training pack. Anchor papers are mid-range exemplars of a score point, while the training pack papers illustrate the range within the score point. CRs working closely with QACs for each content area facilitated the selection process. Finding a sufficient number of papers representing the highest scores is very difficult due to their rarity.

All selected materials were subsequently reviewed by the content representatives from each state. Based

on their recommendations, the anchor exemplars and training packs were modified, finalized, and approved for scorer training.

SELECTING AND TRAINING QUALITY ASSURANCE COORDINATORS AND SENIOR READERS

Because "read-behinds" would be performed by the QACs and SRs in order to moderate the scoring process and maintain the integrity of scores, scoring accuracy was a strong criterion for selecting individuals to fill those positions. Since QACs train readers to score items in particular content areas, they were selected based also on their ability to instruct and on their content area level of expertise. QACs typically are retired teachers. The ratio of QACs and SRs to readers was approximately 1:11.

SELECTING READERS

Reader applicants were required to demonstrate their ability by participating in a preliminary scoring evaluation. The *iScore* system enables Measured Progress to efficiently measure a prospective reader's ability to score student responses accurately. After participating in a training session, applicants are required to achieve at least eighty percent exact scoring agreement for reading and mathematics, seventy percent exact agreement for writing, on a qualifying pack consisting of ten responses to a predetermined item in their content area (or twenty responses in the case of equating items). The qualifying responses are randomly selected from a bank of approximately 150, all of which are selected by QACs and approved by the CRs, developers, and content representatives from each state.

TRAINING READERS

To train readers, QACs demonstrated how to apply the language of the scoring guide to an item's anchor pack exemplars. At the conclusion of anchor pack discussion, readers scored the training pack exemplars. QACs then reviewed the training-pack scoring by the readers and answered any questions readers had.

The optimum ratio of training to scoring hours was determined for divvying readers into content area groups trained to score different items. The resulting amount of time a reader scored a given item was thereby kept short enough to minimize "drift" but long enough to analyze the reader's scoring trends. This scheme helped reconcile the need to provide cost-effective scoring while ensuring that

readers maintain or exceed quality standards.

MONITORING READERS

Training and hand-scoring took place over a period of approximately three weeks. Responses were randomly assigned to readers; thus, each item in a student's response booklet was more than likely scored by a different reader. By using the maximum possible number of readers for each student, the procedure effectively minimized error variance due to reader sampling.

After a reader scored a student response, *iScore* determined whether that response should be scored by a second reader, scored by a QAC or SR, or routed for special attention. QACs and SRs used *iScore* to produce daily reader accuracy and speed reports. They were also able to obtain current reader accuracy speed reports on-line at any time. All common and matrix CR items in reading and mathematics were scored once with a two-percent double-blind (scored independently by two readers) to ensure consistency among readers and accuracy of individual readers. At grades 5 and 8, the common writing prompt was 100% double-blind scored with the requirement that the two scores for each writing component had to be at least adjacent. Non-adjacent scores were arbitrated. The combined scores given by the two readers resulted in the student's raw score on the writing prompt. Each of the three writing CR items was scored once with a two-percent read-behind, and these points were added to the points earned on the writing prompt and the points earned on the ten MC items covering the structures of language and conventions, resulting in the total raw score for writing.

Tables 4-2 and 4-3 present the weighted averages of exact, adjacent, and total percentages of agreement. The weighting was based on the number of responses that were re-scored for each question. (Note: These data underestimate scorer accuracy.) Blanks were included in both read-behind and double-blind scoring. Readers were instructed to score as a zero any questions for which the student had made a mark of any kind. However, in many instances it was impossible for the reader to tell whether a mark on the page was written by the student or whether there was a crease in the paper, bleed-through from the other side of the page, or dust on the image screen. In such instances, these responses were

counted as neither exact nor adjacent agreement, though the effect of blanks and zeroes on student scores was identical.

Table 4-2. 2006-07 NECAP: Percentage Scoring Consistency and Reliability Double-Blind.

Grade	Math			Reading			Writing		
	Exact[1]	Adjacent[1]	Total[1]	Exact	Adjacent	Total	Exact	Adjacent	Total
3	96.3	1.6	97.9	86.9	6.3	93.2			
4	96.6	1.7	98.3	87.5	10.2	97.6			
5	94.1	3.1	97.2	81.6	13.2	94.8	53.5	38.6	92.0
6	93.6	2.9	96.5	82.3	14.2	96.5			
7	93.9	2.8	96.7	81.4	12.2	93.6			
8	95.8	1.9	97.7	83.6	12.7	96.3	57.8	36.4	94.2

[1]Exact = two readers assigned the same score; Adjacent = two readers differed by one point; Total = Exact or adjacent

Table 4-3. 2006-07 NECAP: Percentage Scoring Read-Behind.

Grade	Math			Reading			Writing		
	Exact[1]	Adjacent[1]	Total[1]	Exact	Adjacent	Total	Exact	Adjacent	Total
3	93.8	5.6	99.5	78.6	17.8	96.5			
4	94.9	4.8	99.7	76.4	21.9	98.3			
5	89.4	9.2	98.6	67.9	30.2	98.1	64.3	32.4	96.7
6	88.7	10.0	98.7	70.2	28.3	98.4			
7	87.0	11.4	98.3	68.7	29.8	98.5			
8	90.8	8.2	99.1	71.9	26.6	98.4	64.3	32.4	96.7

[1]Exact = two readers assigned the same score; Adjacent = two readers differed by one point; Total = Exact or adjacent

SCORING LOCATIONS

All of the oversight and administrative controls applied to the *iScore* database were managed for scoring at Measured Progress headquarters in Dover, NH. However, student responses were scored in three locations: Dover, NH; Troy, NY; and Longmont, CO. Table 4-4 shows the locations where all content area/grade level combinations were scored. It is important to note that no single item was scored in more than one location. The *iScore* system monitored accuracy, reliability, and consistency across all scoring locations. Constant communication and coordination were accomplished through e-mail, telephone, faxes, and secure Web sites, to ensure that critical information and scoring modifications were shared/implemented across all scoring locations.

Table 4-4. Content Area/Grade Level Scoring Locations.

Content Area/Grade Level	Dover, NH	Troy, NY	Longmont, CO
Reading Grade 3	X		
Reading Grade 4		X	
Reading Grade 5	X		
Reading Grade 6		X	
Reading Grade 7	X		
Reading Grade 8		X	
Mathematics Grade 3			X
Mathematics Grade 4			X
Mathematics Grade 5			X
Mathematics Grade 6			X
Mathematics Grade 7			X
Mathematics Grade 8			X
Writing Grade 5			X
Writing Grade 8			X

EXTERNAL OBSERVATIONS

The Dover, NH and Longmont, CO scoring locations were visited by at least one representative from each of the three Departments of Education during scoring. State test directors and content specialists from the three states were present at some point at each of the locations during benchmarking, training, and live scoring throughout the scoring window. The state test directors and content specialists from the three states met with program management and scoring management staff from Measured Progress to share their observations and provide feedback. Recommendations that were a result of that meeting will be applied to the next round of scoring in 2007–08.

CHAPTER 5—SCALING AND EQUATING

5.1 ITEM RESPONSE THEORY SCALING

All NECAP items were calibrated using Item Response Theory (IRT). IRT uses mathematical models to define a relationship between an unobserved measure of student performance, usually referred to as theta (θ), and the probability (p) of getting a dichotomous item correct or of getting a particular score on a polytomous item. In IRT, it is assumed that all items are independent measures of the same construct (i.e., of the same θ). Another way to think of θ is as a mathematical representation of the latent trait of interest. Several common IRT models are used to specify the relationship between θ and p (Hambleton and van der Linden, 1997; Hambleton and Swaminathan, 1985). The process of determining the specific mathematical relationship between θ and p is called item calibration. After items are calibrated, they are defined by a set of parameters that specify a nonlinear, monotonically increasing relationship between θ and p. Once the item parameters are known, $\hat{\theta}$, an estimate of θ for each student, can be calculated. ($\hat{\theta}$ is considered to be an estimate of the student's true score or a general representation of student performance. It has characteristics that may make its use preferable to the use of raw scores in equating.)

For NECAP 2006-07, the three-parameter logistic (3PL) model was used for dichotomous items (MC and SA) and the graded-response model (GRM) was used for polytomous items. The 3PL model for dichotomous items can be defined as:

$$P_i\left(1\middle|\theta_j\right) = c_i + \left(1-c_i\right)\frac{\exp Da_i\left(\theta_j - b_i\right)}{1+\exp Da_i\left(\theta_j - b_i\right)}$$

where *i* indexes the items,

j indexes students,
a represents the item discrimination parameter,
b represents the item difficulty parameter,
c is the pseudo-guessing parameter (fixed at 0 for short answer items), and
D is a normalizing constant equal to approximately 1.701.

In the GRM for polytomous items, an item is scored in k+1 graded categories that can be viewed as a set of k dichotomies. At each point of dichotomization (i.e., at each threshold), a two-parameter model can be used. This implies that a polytomous item with k+1 categories can be characterized by k item category threshold curves (ICTC) of the two-parameter logistic form:

$$P_{ik}^{*}\left(1\middle|\theta_j\right)=\frac{\exp Da_i\left(\theta_j-b_i+d_{ik}\right)}{1+\exp Da_i\left(\theta_j-b_i+d_{ik}\right)}$$

where *i* indexes the items,
j indexes students,
k indexes thresholds,
a represents the item discrimination parameter,
b represents the item difficulty parameter,
d represents a category step parameter, and
D is a normalizing constant equal to approximately 1.701.

After computing k item category threshold curves in the GRM, k+1 item category characteristic curves (ICCC) are derived by subtracting adjacent ICTC curves:

$$P_{ik}(1\mid\theta_j)=P_{i(k-1)}{}^{*}(1\mid\theta_j)-P_{ik}{}^{*}(1\mid\theta_j)$$

where P_{ik} represents the probability that the score on item *i* falls in category *k*, and

P_{ik}^{*} represents the probability that the score on item *i* falls above the threshold *k*

($P_{i0}^{*}=1$ and $P_{i(k+1)}^{*}=0$).

The GRM is also commonly expressed as:

$$P_{ik}\left(k\left|\theta_j,\xi_i\right.\right)=\frac{\exp\left[Da_i\left(\theta_j-b_i+d_k\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i+d_k\right)\right]}-\frac{\exp\left[Da_i\left(\theta_j-b_i+d_{k+1}\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i+d_{k+1}\right)\right]}$$

where ξ_i represents the set of item parameters for item *i*.

Finally, the ICC for polytomous items is computed as a weighted sum of ICCCs, where each ICCC is weighted by the score assigned to a corresponding category.

$$P_i(1\,|\,\theta_j)=\sum_k^{m+1} w_{ik}P_{ik}(1\,|\,\theta_j)$$

For more information about item calibration and determination, the reader is referred to Lord and Novick (1968) or Hambleton and Swaminathan (1985).

5.2 EQUATING

The purpose of equating is to ensure that scores obtained from different forms of a test are equivalent to each other. Equating may be used if multiple test forms are administered in the same year, as well as to equate one year's forms to those given in the previous year. Equating ensures that students are not given an unfair advantage or disadvantage because the test form they took is easier or harder than those taken by other students.

The 2006-07 administration of NECAP used a raw score-to-theta equating procedure in which test forms are equated every year to the theta scale of the reference test forms. This is established through the chained linking design, which means that every new form is equated back to the theta scale of the previous year's test form. Since the chain originates from the reference form, it can be assumed that the theta scale of every new test form is the same as the theta scale of the reference form – in the current case, the theta scale of the 2005-06 NECAP

Equating for NECAP used the *anchor-test-nonequivalent-groups* design described by Petersen, Kolen, & Hoover (1989). In this equating design, no assumption is made about the equivalence of the examinee groups taking different test forms (that is, naturally occurring groups are assumed).

Comparability is instead evaluated through utilizing a set of anchor items (i.e., equating items). The NECAP uses an *external* anchor test design, which means that the equating items are not counted toward students' test scores. However, the equating items are designed to mirror the common test in terms of item types and distribution of emphasis. The set of equating items is matrixed across the forms of the test.

Item parameter estimates for 2006-07 were placed on the 2005-06 scale by using the method of Stocking and Lord (1983), which is based on the IRT principle of item parameter invariance. According to this principle, the equating items for both the 2005-06 and 2006-07 NECAP tests should have the same item parameters. The equating procedure was as follows: PARSCALE was used to estimate item parameters for 2006-07 NECAP Math and Reading (the three-parameter logistic model [3PL] for dichotomous items and the graded response model [GRM] for polytomous items). The Stocking and Lord method was employed to find the linear transformation (slope and intercept) that adjusted the equating items' parameter estimates such that the test characteristic curve (TCC; see section 6.5 for a definition of TCCs) was as close as possible to the TCC based on the 2005-06 equating item parameter estimates. (These transformation constants, used to transform all the PARSCALE item parameter estimates from the 2006-07 administration, are given in Appendix C.) Note: Writing was excepted from this equating process; the 2006-07 NECAP writing test forms were pre-equated based on pilot testing in 2004-05 (see the 2005-06 NECAP Technical Report for more details on the pilot). However, the same IRT models as used in all other grade/contents were used for writing (i.e., 3PL and GRM). The final item parameter estimates for all grades and content areas are provided in Appendix G.

Students who took the equating items on the 2006-07 and 2005-06 NECAP tests are not equivalent groups. Item Response Theory (IRT) is particularly useful for equating in nonequivalent group scenarios (Allen & Yen, 1979). The next administration of NECAP, 2007-08, will be scaled to the 2006-07 administration by the same equating method described above.

The Equating Report was submitted to the NECAP state testing directors for their approval prior to production of student reports. The Equating Report is included as Appendix C, and results are discussed more fully in Section 6.7.

5.3 REPORTED SCALE SCORES

DESCRIPTION OF SCALE

A scale was developed for reporting purposes for each NECAP test. These reporting scales are simple linear transformations of the underlying scale (θ) used in the IRT calibrations. The scales were developed such that they ranged from X00 through X80, where X is grade level. In other words, grade 3 scaled scores ranged from 300 to 380, grade 4 from 400 through 480, and so forth through grade 8, where scores ranged from 800 through 880. The lowest scaled score in the *Proficient* range was set at "X40" for each grade level. For example, to be classified in the *Proficient* achievement level or above, a minimum scaled score of 340 was required at grade 3, 440 at grade 4, and so forth.

Scaled scores supplement achievement-level results by providing information that is more specific about the position of a student's results within an achievement level. School- and district-level scaled scores are calculated by computing the average of student-level scaled scores. Students' raw scores (i.e., total number of points) on the 2006-07 NECAP tests were translated to scaled scores using a data analysis process called *scaling*. Scaling simply converts raw points from one scale to another through the TCC. In the same way that a given temperature can be expressed on either Fahrenheit or Celsius scales, or the same distance can be expressed in either miles or kilometers, student scores on the 2006-07 NECAP tests can be expressed in raw or scaled scores.

It is important to note that converting from raw scores to scaled scores does not change students' achievement-level classifications. Given the relative simplicity of raw scores, it is fair to question why scaled scores for NECAP are reported instead of raw scores. Scaled scores simplify the reporting of results across content areas and across successive years. To illustrate, standard-setting typically results

in different *raw* cutscores across content areas.

The raw cut score between *Partially Proficient* and *Proficient* could be, for example, 35 in mathematics but 33 in reading. Both of these raw scores would be transformed to scaled scores of X40, i.e., in the *Proficient* achievement level, just beyond the range of scores associated with the *Partially Proficient* level, as noted above. The same would hold regardless of content area or grade, so one sees that scaled scores facilitate understanding how a student performed. Another advantage of scaled scores comes from their being *linear* transformations of θ. Since the θ scale is used for equating, scaled scores are comparable from one year to the next. Raw scores are not.

CALCULATIONS

The scaled scores are obtained by a simple translation of ability estimates ($\hat{\theta}$) using the linear relationship between threshold values on the θ metric and their equivalent values on the scaled score metric. Students' ability estimates are based on their raw scores and are found by mapping through the TCC. Scaled scores are calculated using the linear equation

$$SS = m\hat{\theta} + b$$

where *m* is the slope and *b* is the intercept. A separate linear transformation was used for each grade/content combination. For the 2006-07 NECAP, each line was determined by fixing both the *Partially Proficient/Proficient* cutscore and the bottom of the scale; that is, the X40 value (e.g., 340 for grade 3) and the X00 value (e.g., 300 for grade 3). The latter was a location on the θ scale beyond the scaling of all the items across the various grade/content combinations. To determine this location, a chance score (approximately equal to a student's expected performance by guessing) was mapped to a value of –4.0 on the θ scale. A raw score of 0 was also assigned a scaled score of X00. The maximum raw score was assigned a scaled score of X80 (e.g., 380 in the case of grade 3).

Because only two points within the θ scaled-score space were fixed, the cutscores between *Substantially Below Proficient* and *Partially Proficient* (SBP/PP) and between *Proficient* and *Proficient with Distinction* (P/PWD) varied across the grade/content combinations.

Table 5-1 represents the scaled cutscores for each grade/content combination (i.e., the minimum scaled score for getting into the next achievement level). It is important to note that the values in Table 5-1 will not change from year to year because the cutscores along the θ scale will not change. In any given year, it may not be possible to attain a particular scaled score, but the scaled score cuts will remain the same.

Table 5-1. NECAP Cut Scores for Each Achievement Level by Grade and Content Area.

Grade	Content	Scale Score Cuts				
		Min	SBP/PP	PP/P	P/PWD	Max
3	Math	300	332	340	353	380
4		400	431	440	455	480
5		500	533	540	554	580
6		600	633	640	653	680
7		700	734	740	752	780
8		800	834	840	852	880
3	Reading	300	331	340	357	380
4		400	431	440	456	480
5		500	530	540	556	580
6		600	629	640	659	680
7		700	729	740	760	780
8		800	828	840	859	880
5	Writing	500	528	540	555	580
8		800	829	840	857	880

SBP = Substantially Below Proficient; PP = Partially Proficient; P = Proficient; PWD = Proficient with Distinction

Table 5-2 shows the cutscores on the θ metric resulting from standard setting (see the 2005-06 NECAP Technical Report for a description of the standard-setting process) and the slope and intercept terms used to calculate the scaled scores. Note that no number in Table 5-2 will change unless the standards are reset.

Table 5-2. NECAP Cutscores (on θ Metric), Intercept, and Slope by Grade and Content Area.

Grade	Content	θ Cuts			Intercept	Slope
		SBP/PP	PP/P	P/PWD		
3	Math	–1.0381	–0.2685	0.9704	342.8782	10.7195
4		–1.1504	–0.37785	0.9493	444.1727	11.0432
5		–0.9279	–0.28455	1.0313	543.0634	10.7659
6		–0.87425	–0.22365	1.03425	642.3690	10.5922
7		–0.70795	–0.0787	1.09945	740.8028	10.2007
8		–0.6444	–0.0286	1.11775	840.2881	10.0720
3	Reading	–1.32285	–0.497	1.0307	345.6751	11.4188
4		–1.173	–0.3142	1.14725	443.4098	10.8525
5		–1.33545	–0.4276	1.04035	544.7878	11.1970
6		–1.47795	–0.51795	1.12545	645.9499	11.4875
7		–1.4833	–0.5223	1.20575	746.0074	11.5019
8		–1.52505	–0.5224	1.1344	846.0087	11.5022
5	Writing	–1.2008	–0.0232	1.5163	540.2334	10.0583
8		–1.0674	–0.0914	1.823	839.1064	9.7766

SBP = Substantially Below Proficient; PP = Partially Proficient; P = Proficient; PWD = Proficient with Distinction

Appendix D contains the raw score—to—scaled score conversion tables. These are the actual tables that were used to determine student scaled scores (along with error bands) and achievement levels.

DISTRIBUTIONS

Appendix E contains the scaled score cumulative density functions. These distributions were calculated using the sparse data matrix files that were used in the IRT calibrations. For each grade/content, these distributions show the cumulative percentage of students scoring at or below a particular scaled score across the entire scaled score range.

SECTION II
STATISTICAL AND PSYCHOMETRIC SUMMARIES

CHAPTER 6—ITEM ANALYSES

As noted in Brown (1983), "A test is only as good as the items it contains." A complete evaluation of a test's quality must include an evaluation of each question. Both the *Standards for Educational and Psychological Testing* (AERA, 1999) and the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988) include standards for identifying quality questions. Questions should assess only knowledge or skills that are identified as part of the domain being measured and should avoid assessing irrelevant factors. They should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. Further, questions must not unfairly disadvantage test takers from particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses were conducted to ensure that NECAP questions met these standards. Qualitative analyses were discussed in Chapter 2 ("Development and Test Design"). The following discussion focuses on several categories of quantitative evaluation of 2006-07 NECAP items: (a) difficulty indices, (b) item-test correlations, (c) subgroup differences in item performance (differential item functioning), (d) dimensionality analyses, (e) IRT analyses, and (f) equating results.

6.1 DIFFICULTY INDICES

All 2006-07 NECAP items were evaluated in terms of difficulty according to standard classical test theory (CTT) practice. The expected item difficulty, also known as the *p-value*, is the main index of item difficulty under the CTT framework. This index measures an item's difficulty by averaging the proportion of points received across all students who took the item. MC items were scored

dichotomously (correct vs. incorrect), so for these items, the difficulty index is simply the proportion of students who correctly answered the item. To place all item types on the same 0–1 scale, the p-value of an OR item was computed as the average score on the item divided by its maximum possible score. Although the p-value is traditionally called a measure of difficulty, it is properly interpreted as an *easiness* index, because larger values indicate easier items. An index of 0 indicates that no student received credit for the item. At the opposite extreme, an index of 1 indicates that every student received full credit for the item.

Items that are answered correctly by almost all students provide little information about differences in student ability, but they do indicate knowledge or skills that have been mastered by most students. The converse is true of items that are incorrectly answered by most students. In general, to provide the most precise measurement, difficulty indices should range from near-chance performance (0.25 for four-option MC items, 0.00 for CR items) to 0.90. Experience has indicated that items conforming to this guideline tend to provide satisfactory statistical information for the bulk of the student population. However, on a criterion-referenced test such as NECAP, it may be appropriate to include some items with difficulty values outside this region in order to measure well, throughout the range, the skill present at a given grade. Having a range of item difficulties also helps to ensure that the test does not exhibit an excess of scores at the floor or ceiling of the distribution.

6.2 ITEM–TEST CORRELATIONS

It is a desirable feature of an item when higher-ability students perform better on it than do lower-ability students. A commonly used measure of this characteristic is the correlation between total test score and student performance on the item. Within CTT, this item-test correlation is referred to as the item's *discrimination*, because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. For polytomous items on the 2006-07 NECAP, the *Pearson product-moment correlation* was used as the item discrimination index and the *point-biserial correlation* was used for dichotomous items.

The theoretical range of these statistics is –1.0 to +1.0, with a typical range from +0.2 to +0.6.

One can think of a discrimination index as a measure of how closely an item assesses the same knowledge and skills as other items that contribute to the criterion total score; in other words, the discrimination index can be interpreted as a measure of construct consistency. In light of this, it is quite important that an appropriate total score criterion be selected. For the 2006-07 NECAP, raw score – the sum of student scores on the common items – was selected. Item-test correlations were computed for each common item, and results are summarized in the next section.

6.3 SUMMARY OF ITEM ANALYSIS RESULTS

Summary statistics of the difficulty and discrimination indices by grade and content area are provided in Appendix F. Table F-1 displays the means and standard deviations of p-values and discriminations by form for each grade and content area of the 2006-07 NECAP administration. p-value means ranged between 0.42 and 0.75, and their standard deviations ranged between 0.11 and 0.25 across all grades, subject areas, and forms. Discrimination (item-total correlation) means ranged between 0.37 and 0.52, standard deviations between 0.04 and 0.19.

Table F-2 presents summary statistics (means and standard deviations) for the p-values and discriminations by item type (MC and OR) and aggregated over both item types. Across all grades and content areas, mean p-values for MC items fell between 0.53 and 0.80, for OR items between 0.34 and 0.71, and for both item types together between 0.46 and 0.75. Mean discrimination indices for MC items ranged between 0.34 and 0.44, for OR items between 0.44 and 0.65, and for all items together between 0.38 and 0.47.

Finally, Table F-3 shows the number, relative percentages, and cumulative percentages of common items that had difficulty or discrimination values within stated ranges. p-values and discrimination indices were generally in expected ranges. Very few items were answered correctly at near-chance or near-perfect rates, and positive discrimination indices indicate that students who performed well on individual items tended to perform well overall. Though it is not inappropriate to

include low discriminating items or very difficult or very easy items, to ensure that the entire ability spectrum is appropriately covered, there were very few such items on the NECAP tests.

A comparison of indices across grade levels is complicated because these indices are population-dependent. Direct comparisons would require that either the items or students were common across groups. As that was not the case, it cannot be determined whether differences in item functioning across grade levels were due to differences in student cohorts' abilities or differences in item-set difficulties or both. However, one noteworthy statistical trend in math was that p-values tended to be highest at the lower grades.

Comparing the difficulty indices between MC and OR items is also inappropriate. MC items can be answered correctly by guessing; thus, it is not surprising that the p-values for MC items were higher than those for OR items. Similarly, because of partial-credit scoring, the discrimination indices of OR items tended to be larger than those of MC items.

6.4 DIFFERENTIAL ITEM FUNCTIONING

The *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988) explicitly states that subgroup differences in performance should be examined when sample sizes permit, and actions should be taken to make certain that differences in performance are due to construct-relevant, rather than construct-irrelevant, factors. The *Standards for Educational and Psychological Testing* (AERA, 1999) includes similar guidelines. As part of the effort to identify such problems, 2006-07 NECAP items were evaluated by means of DIF statistics.

DIF procedures are designed to identify items on which the performance by certain subgroups of interest differs after controlling for construct-relevant achievement. For the 2006-07 NECAP, the standardization DIF procedure (Dorans & Kulick, 1986) was employed. This procedure calculates the difference in item performance for two groups of students (at a time) matched for achievement on the total test. Specifically, average item performance is calculated for students at every total score. Then an overall average is calculated, weighting the total score distribution so that it is the same for the two

groups. The criterion (matching) score for 2006-07 NECAP was computed two ways. For common items, total score was the sum of scores on common items. The total score criterion for matrix items was the sum of item scores on both common and matrix items (excluding field-test items). Based on experience, this dual definition of criterion scores has worked well in identifying problematic common and matrix items.

Differential performances between groups may or may not be indicative of bias in the test. Group differences in course-taking patterns, interests, or school curricula can lead to DIF. If subgroup differences are related to construct-relevant factors, items should be considered for inclusion on a test.

Computed DIF indices have a theoretical range from –1.00 to 1.00 for MC items; those for OR items are adjusted to the same scale. For reporting purposes, items were categorized according to DIF index range guidelines suggested by Dorans and Holland (1993). Indices between –0.05 and 0.05 (Type A) can be considered "negligible." Most items should fall in this range. DIF indices between –0.10 and –0.05 or between 0.05 and 0.10 (Type B) can be considered "low DIF" but should be inspected to ensure that no possible effect is overlooked. Items with DIF indices outside the [–0.10, 0.10] range (Type C) can be considered "high DIF" and should trigger careful test.

The following series of three tables presents the number of 2006-07 NECAP items classified into each DIF category, broken down by grade, subject area form, and item type. Results are given, respectively, for comparisons between Male and Female, White and Black, and White and Hispanic. Note that "Form 00" contains the common items that are used in calculating reported scores for students. In addition to the DIF categories defined above (i.e., Types A, B, and C), "Type D" in the tables indicates that there were not enough students in the grouping to perform a reliable DIF analysis (i.e., fewer than 200 in at least one of the subgroups).

Table 6-1. Number of NECAP 2006-07 Items Classified into Differential Item Functioning (DIF) Categories by Grade, Subject, and Test Form: Male versus Female.

Grade	Subject	Form	All A	All B	All C	All D	MC A	MC B	MC C	MC D	OR A	OR B	OR C	OR D
3	Math	00	53	2	0	0	34	1	0	0	19	1	0	0
		01	10	0	0	0	6	0	0	0	4	0	0	0
		02	10	0	0	0	6	0	0	0	4	0	0	0
		03	9	1	0	0	5	1	0	0	4	0	0	0
		04	9	1	0	0	5	1	0	0	4	0	0	0
		05	10	0	0	0	6	0	0	0	4	0	0	0
		06	9	1	0	0	5	1	0	0	4	0	0	0
		07	9	1	0	0	5	1	0	0	4	0	0	0
		08	10	0	0	0	6	0	0	0	4	0	0	0
		09	10	0	0	0	6	0	0	0	4	0	0	0
	Reading	00	34	0	0	0	28	0	0	0	6	0	0	0
		01	16	1	0	0	14	0	0	0	2	1	0	0
		02	15	2	0	0	13	1	0	0	2	1	0	0
		03	16	1	0	0	14	0	0	0	2	1	0	0
4	Math	00	50	3	2	0	31	3	1	0	19	0	1	0
		01	9	1	0	0	5	1	0	0	4	0	0	0
		02	10	0	0	0	6	0	0	0	4	0	0	0
		03	10	0	0	0	6	0	0	0	4	0	0	0
		04	9	1	0	0	5	1	0	0	4	0	0	0
		05	10	0	0	0	6	0	0	0	4	0	0	0
		06	9	1	0	0	5	1	0	0	4	0	0	0
		07	9	1	0	0	6	0	0	0	3	1	0	0
		08	10	0	0	0	6	0	0	0	4	0	0	0
		09	10	0	0	0	6	0	0	0	4	0	0	0
	Reading	00	32	2	0	0	27	1	0	0	5	1	0	0
		01	16	0	1	0	13	0	1	0	3	0	0	0
		02	16	1	0	0	13	1	0	0	3	0	0	0
		03	13	4	0	0	11	3	0	0	2	1	0	0
5	Math	00	45	3	0	0	30	2	0	0	15	1	0	0
		01	11	0	0	0	6	0	0	0	5	0	0	0
		02	11	0	0	0	6	0	0	0	5	0	0	0
		03	9	2	0	0	6	0	0	0	3	2	0	0
		04	11	0	0	0	6	0	0	0	5	0	0	0
		05	11	0	0	0	6	0	0	0	5	0	0	0
		06	11	0	0	0	6	0	0	0	5	0	0	0
		07	11	0	0	0	6	0	0	0	5	0	0	0
		08	11	0	0	0	6	0	0	0	5	0	0	0
		09	9	2	0	0	5	1	0	0	4	1	0	0
	Reading	00	29	3	2	0	24	2	2	0	5	1	0	0
		01	14	3	0	0	11	3	0	0	3	0	0	0
		02	16	1	0	0	13	1	0	0	3	0	0	0
		03	17	0	0	0	14	0	0	0	3	0	0	0
	Writing	01	17	0	0	0	10	0	0	0	7	0	0	0

cont'd

Table 6-1. Number of NECAP 2006-07 Items Classified into Differential Item Functioning (DIF) Categories by Grade, Subject, and Test Form: Male versus Female.

Grade	Subject	Form	All A	All B	All C	All D	MC A	MC B	MC C	MC D	OR A	OR B	OR C	OR D
6	Math	00	42	4	2	0	27	3	2	0	15	1	0	0
		01	8	3	0	0	5	1	0	0	3	2	0	0
		02	11	0	0	0	6	0	0	0	5	0	0	0
		03	8	3	0	0	3	3	0	0	5	0	0	0
		04	10	1	0	0	5	1	0	0	5	0	0	0
		05	10	1	0	0	5	1	0	0	5	0	0	0
		06	10	0	1	0	5	0	1	0	5	0	0	0
		07	9	2	0	0	6	0	0	0	3	2	0	0
		08	11	0	0	0	6	0	0	0	5	0	0	0
		09	8	3	0	0	3	3	0	0	5	0	0	0
	Reading	00	28	4	2	0	24	2	2	0	4	2	0	0
		01	14	2	1	0	12	1	1	0	2	1	0	0
		02	16	1	0	0	14	0	0	0	2	1	0	0
		03	12	4	1	0	10	3	1	0	2	1	0	0
7	Math	00	40	8	0	0	28	4	0	0	12	4	0	0
		01	10	0	1	0	6	0	0	0	4	0	1	0
		02	8	3	0	0	5	1	0	0	3	2	0	0
		03	10	1	0	0	5	1	0	0	5	0	0	0
		04	9	2	0	0	4	2	0	0	5	0	0	0
		05	7	3	1	0	3	2	1	0	4	1	0	0
		06	10	1	0	0	6	0	0	0	4	1	0	0
		07	9	1	1	0	6	0	0	0	3	1	1	0
		08	9	2	0	0	5	1	0	0	4	1	0	0
		09	11	0	0	0	6	0	0	0	5	0	0	0
	Reading	00	26	7	1	0	21	6	1	0	5	1	0	0
		01	16	1	0	0	13	1	0	0	3	0	0	0
		02	12	4	1	0	11	2	1	0	1	2	0	0
		03	13	4	0	0	10	4	0	0	3	0	0	0
8	Math	00	43	5	0	0	29	3	0	0	14	2	0	0
		01	9	2	0	0	6	0	0	0	3	2	0	0
		02	9	2	0	0	4	2	0	0	5	0	0	0
		03	8	3	0	0	5	1	0	0	3	2	0	0
		04	10	1	0	0	5	1	0	0	5	0	0	0
		05	7	4	0	0	3	3	0	0	4	1	0	0
		06	11	0	0	0	6	0	0	0	5	0	0	0
		07	10	1	0	0	6	0	0	0	4	1	0	0
		08	7	4	0	0	3	3	0	0	4	1	0	0
		09	8	3	0	0	4	2	0	0	4	1	0	0
	Reading	00	31	3	0	0	25	3	0	0	6	0	0	0
		01	12	2	3	0	10	1	3	0	2	1	0	0
		02	14	3	0	0	13	1	0	0	1	2	0	0
		03	12	5	0	0	12	2	0	0	0	3	0	0
	Writing	01	17	0	0	0	10	0	0	0	7	0	0	0

All = MC and OR items; **MC** = Multiple-choice items; **OR** = Open-response items;
A = "negligible" DIF; **B** = "low" DIF; **C** = "high" DIF; **D** = not enough students to perform reliable DIF analysis

Table 6-2. Number of NECAP 2006-07 Items Classified into Differential Item Functioning (DIF) Categories by Grade, Subject, and Test Form: White versus Black.

Grade	Subject	Form	All A	All B	All C	All D	MC A	MC B	MC C	MC D	OR A	OR B	OR C	OR D
3	Math	00	48	7	0	0	32	3	0	0	16	4	0	0
		01	0	0	0	10	0	0	0	6	0	0	0	4
		02	0	0	0	10	0	0	0	6	0	0	0	4
		03	0	0	0	10	0	0	0	6	0	0	0	4
		04	0	0	0	10	0	0	0	6	0	0	0	4
		05	0	0	0	10	0	0	0	6	0	0	0	4
		06	0	0	0	10	0	0	0	6	0	0	0	4
		07	0	0	0	10	0	0	0	6	0	0	0	4
		08	0	0	0	10	0	0	0	6	0	0	0	4
		09	0	0	0	10	0	0	0	6	0	0	0	4
	Reading	00	34	0	0	0	28	0	0	0	6	0	0	0
		01	0	0	0	17	0	0	0	14	0	0	0	3
		02	0	0	0	17	0	0	0	14	0	0	0	3
		03	0	0	0	17	0	0	0	14	0	0	0	3
4	Math	00	46	7	2	0	30	3	2	0	16	4	0	0
		01	0	0	0	10	0	0	0	6	0	0	0	4
		02	0	0	0	10	0	0	0	6	0	0	0	4
		03	0	0	0	10	0	0	0	6	0	0	0	4
		04	0	0	0	10	0	0	0	6	0	0	0	4
		05	0	0	0	10	0	0	0	6	0	0	0	4
		06	0	0	0	10	0	0	0	6	0	0	0	4
		07	0	0	0	10	0	0	0	6	0	0	0	4
		08	0	0	0	10	0	0	0	6	0	0	0	4
		09	0	0	0	10	0	0	0	6	0	0	0	4
	Reading	00	29	5	0	0	23	5	0	0	6	0	0	0
		01	0	0	0	17	0	0	0	14	0	0	0	3
		02	0	0	0	17	0	0	0	14	0	0	0	3
		03	0	0	0	17	0	0	0	14	0	0	0	3
5	Math	00	44	4	0	0	30	2	0	0	14	2	0	0
		01	0	0	0	11	0	0	0	6	0	0	0	5
		02	0	0	0	11	0	0	0	6	0	0	0	5
		03	0	0	0	11	0	0	0	6	0	0	0	5
		04	0	0	0	11	0	0	0	6	0	0	0	5
		05	0	0	0	11	0	0	0	6	0	0	0	5
		06	0	0	0	11	0	0	0	6	0	0	0	5
		07	0	0	0	11	0	0	0	6	0	0	0	5
		08	0	0	0	11	0	0	0	6	0	0	0	5
		09	0	0	0	11	0	0	0	6	0	0	0	5
	Reading	00	27	6	1	0	21	6	1	0	6	0	0	0
		01	0	0	0	17	0	0	0	14	0	0	0	3
		02	0	0	0	17	0	0	0	14	0	0	0	3
		03	0	0	0	17	0	0	0	14	0	0	0	3
	Writing	01	15	1	1	0	8	1	1	0	7	0	0	0

cont'd

Table 6-2. Number of NECAP 2006-07 Items Classified into Differential Item Functioning (DIF) Categories by Grade, Subject, and Test Form: White versus Black.

Grade	Subject	Form	All A	All B	All C	All D	MC A	MC B	MC C	MC D	OR A	OR B	OR C	OR D
6	Math	00	44	4	0	0	28	4	0	0	16	0	0	0
		01	0	0	0	11	0	0	0	6	0	0	0	5
		02	0	0	0	11	0	0	0	6	0	0	0	5
		03	0	0	0	11	0	0	0	6	0	0	0	5
		04	0	0	0	11	0	0	0	6	0	0	0	5
		05	0	0	0	11	0	0	0	6	0	0	0	5
		06	0	0	0	11	0	0	0	6	0	0	0	5
		07	0	0	0	11	0	0	0	6	0	0	0	5
		08	0	0	0	11	0	0	0	6	0	0	0	5
		09	0	0	0	11	0	0	0	6	0	0	0	5
	Reading	00	26	7	1	0	20	7	1	0	6	0	0	0
		01	0	0	0	17	0	0	0	14	0	0	0	3
		02	0	0	0	17	0	0	0	14	0	0	0	3
		03	0	0	0	17	0	0	0	14	0	0	0	3
7	Math	00	45	1	2	0	30	1	1	0	15	0	1	0
		01	0	0	0	11	0	0	0	6	0	0	0	5
		02	0	0	0	11	0	0	0	6	0	0	0	5
		03	0	0	0	11	0	0	0	6	0	0	0	5
		04	0	0	0	11	0	0	0	6	0	0	0	5
		05	0	0	0	11	0	0	0	6	0	0	0	5
		06	0	0	0	11	0	0	0	6	0	0	0	5
		07	0	0	0	11	0	0	0	6	0	0	0	5
		08	0	0	0	11	0	0	0	6	0	0	0	5
		09	0	0	0	11	0	0	0	6	0	0	0	5
	Reading	00	31	2	1	0	25	2	1	0	6	0	0	0
		01	0	0	0	17	0	0	0	14	0	0	0	3
		02	0	0	0	17	0	0	0	14	0	0	0	3
		03	0	0	0	17	0	0	0	14	0	0	0	3
8	Math	00	39	9	0	0	25	7	0	0	14	2	0	0
		01	0	0	0	11	0	0	0	6	0	0	0	5
		02	0	0	0	11	0	0	0	6	0	0	0	5
		03	0	0	0	11	0	0	0	6	0	0	0	5
		04	0	0	0	11	0	0	0	6	0	0	0	5
		05	0	0	0	11	0	0	0	6	0	0	0	5
		06	0	0	0	11	0	0	0	6	0	0	0	5
		07	0	0	0	11	0	0	0	6	0	0	0	5
		08	0	0	0	11	0	0	0	6	0	0	0	5
		09	0	0	0	11	0	0	0	6	0	0	0	5
	Reading	00	29	4	1	0	23	4	1	0	6	0	0	0
		01	0	0	0	17	0	0	0	14	0	0	0	3
		02	0	0	0	17	0	0	0	14	0	0	0	3
		03	0	0	0	17	0	0	0	14	0	0	0	3
	Writing	01	15	2	0	0	8	2	0	0	7	0	0	0

All = MC and OR items; **MC** = Multiple-choice items; **OR** = Open-response items;
A = "negligible" DIF; **B** = "low" DIF; **C** = "high" DIF; **D** = not enough students to perform reliable DIF analysis

Table 6-3. Number of NECAP 2006-07 Items Classified into Differential Item Functioning (DIF) Categories by Grade, Subject, and Test Form: White versus Hispanic.

Grade	Subject	Form	All A	All B	All C	All D	MC A	MC B	MC C	MC D	OR A	OR B	OR C	OR D
3	Math	00	41	14	0	0	26	9	0	0	15	5	0	0
		01	5	4	1	0	2	3	1	0	3	1	0	0
		02	4	5	1	0	3	2	1	0	1	3	0	0
		03	8	2	0	0	4	2	0	0	4	0	0	0
		04	7	2	1	0	6	0	0	0	1	2	1	0
		05	9	0	1	0	5	0	1	0	4	0	0	0
		06	9	1	0	0	5	1	0	0	4	0	0	0
		07	6	3	1	0	5	0	1	0	1	3	0	0
		08	6	3	1	0	4	2	0	0	2	1	1	0
		09	9	1	0	0	6	0	0	0	3	1	0	0
	Reading	00	32	2	0	0	28	0	0	0	4	2	0	0
		01	14	2	1	0	11	2	1	0	3	0	0	0
		02	14	2	1	0	11	2	1	0	3	0	0	0
		03	16	1	0	0	13	1	0	0	3	0	0	0
4	Math	00	44	10	1	0	30	4	1	0	14	6	0	0
		01	6	3	1	0	4	1	1	0	2	2	0	0
		02	7	3	0	0	6	0	0	0	1	3	0	0
		03	8	2	0	0	5	1	0	0	3	1	0	0
		04	8	0	2	0	4	0	2	0	4	0	0	0
		05	6	2	2	0	5	1	0	0	1	1	2	0
		06	6	3	1	0	4	2	0	0	2	1	1	0
		07	6	2	2	0	4	1	1	0	2	1	1	0
		08	8	2	0	0	5	1	0	0	3	1	0	0
		09	7	2	1	0	5	1	0	0	2	1	1	0
	Reading	00	26	7	1	0	20	7	1	0	6	0	0	0
		01	13	3	1	0	11	2	1	0	2	1	0	0
		02	15	2	0	0	12	2	0	0	3	0	0	0
		03	10	5	2	0	8	4	2	0	2	1	0	0
5	Math	00	40	8	0	0	26	6	0	0	14	2	0	0
		01	9	2	0	0	6	0	0	0	3	2	0	0
		02	8	3	0	0	4	2	0	0	4	1	0	0
		03	8	3	0	0	4	2	0	0	4	1	0	0
		04	10	1	0	0	5	1	0	0	5	0	0	0
		05	10	1	0	0	5	1	0	0	5	0	0	0
		06	7	3	1	0	3	2	1	0	4	1	0	0
		07	8	3	0	0	4	2	0	0	4	1	0	0
		08	9	2	0	0	4	2	0	0	5	0	0	0
		09	11	0	0	0	6	0	0	0	5	0	0	0
	Reading	00	22	10	2	0	16	10	2	0	6	0	0	0
		01	12	4	1	0	9	4	1	0	3	0	0	0
		02	10	7	0	0	9	5	0	0	1	2	0	0
		03	10	4	3	0	7	4	3	0	3	0	0	0
	Writing	01	14	3	0	0	7	3	0	0	7	0	0	0

cont'd

Table 6-3. Number of NECAP 2006-07 Items Classified into Differential Item Functioning (DIF) Categories by Grade, Subject, and Test Form: White versus Hispanic.

Grade	Subject	Form	All A	All B	All C	All D	MC A	MC B	MC C	MC D	OR A	OR B	OR C	OR D
6	Math	00	41	6	1	0	26	5	1	0	15	1	0	0
		01	9	2	0	0	5	1	0	0	4	1	0	0
		02	6	4	1	0	4	2	0	0	2	2	1	0
		03	8	2	1	0	4	2	0	0	4	0	1	0
		04	8	2	1	0	3	2	1	0	5	0	0	0
		05	8	2	1	0	5	1	0	0	3	1	1	0
		06	7	3	1	0	4	2	0	0	3	1	1	0
		07	11	0	0	0	6	0	0	0	5	0	0	0
		08	6	4	1	0	4	2	0	0	2	2	1	0
		09	8	2	1	0	4	2	0	0	4	0	1	0
	Reading	00	27	5	2	0	21	5	2	0	6	0	0	0
		01	10	4	3	0	7	4	3	0	3	0	0	0
		02	10	7	0	0	7	7	0	0	3	0	0	0
		03	10	4	3	0	7	4	3	0	3	0	0	0
7	Math	00	42	4	2	0	28	2	2	0	14	2	0	0
		01	10	1	0	0	5	1	0	0	5	0	0	0
		02	8	3	0	0	3	3	0	0	5	0	0	0
		03	10	1	0	0	6	0	0	0	4	1	0	0
		04	9	1	1	0	4	1	1	0	5	0	0	0
		05	8	3	0	0	3	3	0	0	5	0	0	0
		06	9	2	0	0	4	2	0	0	5	0	0	0
		07	8	3	0	0	4	2	0	0	4	1	0	0
		08	8	2	1	0	5	0	1	0	3	2	0	0
		09	11	0	0	0	6	0	0	0	5	0	0	0
	Reading	00	23	10	1	0	17	10	1	0	6	0	0	0
		01	13	2	2	0	10	2	2	0	3	0	0	0
		02	10	5	2	0	8	4	2	0	2	1	0	0
		03	13	3	1	0	10	3	1	0	3	0	0	0
8	Math	00	34	13	1	0	22	9	1	0	12	4	0	0
		01	7	2	2	0	4	1	1	0	3	1	1	0
		02	8	3	0	0	4	2	0	0	4	1	0	0
		03	11	0	0	0	6	0	0	0	5	0	0	0
		04	11	0	0	0	6	0	0	0	5	0	0	0
		05	9	1	1	0	5	0	1	0	4	1	0	0
		06	10	1	0	0	5	1	0	0	5	0	0	0
		07	8	3	0	0	4	2	0	0	4	1	0	0
		08	8	3	0	0	5	1	0	0	3	2	0	0
		09	9	2	0	0	4	2	0	0	5	0	0	0
	Reading	00	26	7	1	0	20	7	1	0	6	0	0	0
		01	10	4	3	0	7	4	3	0	3	0	0	0
		02	10	5	2	0	7	5	2	0	3	0	0	0
		03	11	5	1	0	9	4	1	0	2	1	0	0
	Writing	01	13	4	0	0	6	4	0	0	7	0	0	0

All = MC and OR items; **MC** = Multiple-choice items; **OR** = Open-response items;
A = "negligible" DIF; **B** = "low" DIF; **C** = "high" DIF; **D** = not enough students to perform reliable DIF analysis

The tables show that the majority of DIF distinctions in the 2006-07 NECAP tests were "Type A," i.e., "negligible" DIF (Dorans and Holland , 1993). Although there were items with DIF indices in the "low" or "high" categories, this does not necessarily indicate that the items are biased. Both the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988) and the *Standards for Educational and Psychological Testing* (AERA, 1999) assert that test items must be free from construct-irrelevant sources of differential difficulty. If subgroup differences in performance can be plausibly attributed to construct-relevant factors, the items may be included on a test. What is important is to determine whether the cause of this differential performance is construct-relevant.

Table 6-4 presents the number of items classified into each DIF category by direction, comparing males and females. For example, the "F_A" column denotes the total number of items classified as "negligible" DIF on which females performed better than males relative to performance on the test as a whole. The "M_A" column next to it gives the total number of "negligible" DIF items on which males performed better than females relative to performance on the test as a whole. The "N_A" and "P_A" columns display the aggregate number and proportion of "negligible" DIF items, respectively. To provide a complete summary across items, both common and matrix items are included in the tally that falls into each category. Results are broken out by grade, content area, and item type.

Table 6-4. Number and Proportion of NECAP 2006-07 Items Classified into Each DIF Category and Direction by Item Type: Male versus Female.

Grade	Subject	Item Type	F_A	M_A	N_A	P_A	F_B	M_B	N_B	P_B	F_C	M_C	N_C	P_C	N_D	P_D
3	Math	MC	52	32	84	0.94	2	3	5	0.06	0	0	0	0.00	0	0.00
		OR	32	23	55	0.98	0	1	1	0.02	0	0	0	0.00	0	0.00
	Reading	MC	43	26	69	0.99	0	1	1	0.01	0	0	0	0.00	0	0.00
		OR	8	4	12	0.80	2	1	3	0.20	0	0	0	0.00	0	0.00
4	Math	MC	56	26	82	0.92	1	5	6	0.07	0	1	1	0.01	0	0.00
		OR	39	15	54	0.96	1	0	1	0.02	0	1	1	0.02	0	0.00
	Reading	MC	30	34	64	0.91	0	5	5	0.07	0	1	1	0.01	0	0.00
		OR	9	4	13	0.87	2	0	2	0.13	0	0	0	0.00	0	0.00
5	Math	MC	35	48	83	0.97	1	2	3	0.03	0	0	0	0.00	0	0.00
		OR	32	25	57	0.93	2	2	4	0.07	0	0	0	0.00	0	0.00
	Reading	MC	28	34	62	0.89	0	6	6	0.09	0	2	2	0.03	0	0.00
		OR	14	0	14	0.93	1	0	1	0.07	0	0	0	0.00	0	0.00
	Writing	MC	5	5	10	1.00	0	0	0	0.00	0	0	0	0.00	0	0.00
		OR	7	0	7	1.00	0	0	0	0.00	0	0	0	0.00	0	0.00
6	Math	MC	37	34	71	0.83	2	10	12	0.14	1	2	3	0.03	0	0.00
		OR	34	22	56	0.92	2	3	5	0.08	0	0	0	0.00	0	0.00
	Reading	MC	23	37	60	0.86	0	6	6	0.09	0	4	4	0.06	0	0.00
		OR	10	0	10	0.67	5	0	5	0.33	0	0	0	0.00	0	0.00
7	Math	MC	37	37	74	0.86	1	10	11	0.13	0	1	1	0.01	0	0.00
		OR	22	27	49	0.80	7	3	10	0.16	0	2	2	0.03	0	0.00
	Reading	MC	28	27	55	0.79	1	12	13	0.19	0	2	2	0.03	0	0.00
		OR	12	0	12	0.80	3	0	3	0.20	0	0	0	0.00	0	0.00
8	Math	MC	33	38	71	0.83	4	11	15	0.17	0	0	0	0.00	0	0.00
		OR	31	20	51	0.84	9	1	10	0.16	0	0	0	0.00	0	0.00
	Reading	MC	31	29	60	0.86	0	7	7	0.10	0	3	3	0.04	0	0.00
		OR	9	0	9	0.60	6	0	6	0.40	0	0	0	0.00	0	0.00
	Writing	MC	4	6	10	1.00	0	0	0	0.00	0	0	0	0.00	0	0.00
		OR	7	0	7	1.00	0	0	0	0.00	0	0	0	0.00	0	0.00

F_ = items on which females performed better than males (controlling for total test score); **M_** = items on which males performed better than females, (controlling for total test score); **N_** = number of items; **P_** = proportion of items
_A = "negligible" DIF; **_B** = "low" DIF; **_C** = "high" DIF; **_D** = not enough students to perform a reliable DIF analysis

6.5 DIMENSIONALITY ANALYSES

The DIF analyses of section 6.4 were performed to identify items which showed evidence of differences in performance between pairs of subgroups beyond that which would be expected based on the primary construct that underlies total test score (also known as the "primary dimension;" for example, general achievement in math). When items are flagged for DIF, statistical evidence points to their measuring an additional dimension(s) to the primary dimension.

Because tests are constructed with multiple content area subcategories, and their associated knowledge and skills, the potential exists for a large number of dimensions being invoked beyond the

common primary dimension. Generally, the subcategories are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of the variance in test scores. In fact, the presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional IRT models that are used for calibrating, linking, scaling, and equating the NECAP test forms. As noted in the previous section, a statistically significant DIF result does not automatically imply that an item is measuring an *irrelevant* construct or dimension. An item could be flagged for DIF because it measures one of the construct-*relevant* dimensions of a subcategory's knowledge and skills.

The purpose of dimensionality analysis is to investigate whether violation of the assumption of test unidimensionality is statistically detectable and, if so, (a) the degree to which unidimensionality is violated and (b) the nature of the multidimensionality. Findings from dimensionality (DIM) analyses performed on the 2006-07 NECAP common items for Math, Reading, and Writing are reported below. (Note: only common items were analyzed since they are used for score reporting.)

The DIM analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Both of these methods use as their basic statistical building block the estimated average conditional covariances for item pairs. A conditional covariance is the covariance between two items conditioned on total score for the rest of the test, and the average conditional covariance is obtained by averaging over all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected scores. Non-zero conditional covariances are essentially violations of the principle of local independence, and local *dependence* implies multidimensionality. Thus, non-random patterns of positive and negative conditional covariances are indicative of multidimensionality.

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. The data are first divided into a training sample and a cross-validation sample.

Then an exploratory analysis of the conditional covariances is conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The cross-validation sample is then used to test whether the conditional covariances of the selected cluster of items displays local dependence, conditioning on total score on the non-clustered items. The DIMTEST statistic follows a standard normal distribution under the null hypothesis of unidimensionality.

DETECT is an effect-size measure of multidimensionality. As with DIMTEST, the data are first divided into a training sample and a cross-validation sample. The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative conditional covariances from different clusters. Next, the clusters from the training sample are used with the cross-validation sample data to average the conditional covariances: within-cluster conditional covariances are summed, from this sum the between-cluster conditional covariances are subtracted, this difference is divided by the total number of item pairs, and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality), values of 0.2 to 0.4 weak to moderate multidimensionality; values of 0.4 to 1.0 moderate to strong multidimensionality, and values greater than 1.0 very strong multidimensionality.

DIMTEST and DETECT were applied to the 2006-07 NECAP. The data for each grade and content area were split into a training sample and a cross-validation sample. Every grade/content area combination had at least 32,000 student examinees, so every training sample and cross-validation sample had at least 16,000 students. DIMTEST was then applied to every grade/content area. DETECT was applied to each dataset for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

Because of the very large sample sizes of NECAP, DIMTEST was sensitive even to quite small violations of unidimensionality, and the null hypothesis was strongly rejected for every dataset

($p \leq 0.00005$ in every case). Extremely small effect sizes are not generally indicative of serious multidimensionality, since strict unidimensionality is an idealization that almost never holds exactly for a given dataset. Thus, it was important to use DETECT to estimate the effect size of the violations of local independence found by DIMTEST. Table 6-5 displays the multidimensional effect size estimates from DETECT.

Table 6-5. 2006-07 NECAP: Multidimensionality Effect Sizes by Grade and Subject.

Grade	Subject	Multidimensionality Effect Size
3	Math	0.20
	Reading	0.24
4	Math	0.16
	Reading	0.15
5	Math	0.20
	Reading	0.25
	Writing	0.18
6	Math	0.15
	Reading	0.17
7	Math	0.20
	Reading	0.19
8	Math	0.17
	Reading	0.21
	Writing	0.21

All of the DETECT values indicated very weak to weak multidimensionality. The Reading test forms tended to show slightly greater multidimensionality than did the Math or Writing, but still towards the weak end of the 0.20 to 0.40 range. We also investigated how DETECT divided the tests into clusters to see if there were any discernable patterns with respect to the item types (i.e., MC, SA, and CR). The Math clusters showed no discernable patterns. For Reading and Writing, however, there was a strong tendency for the MC items to cluster separately from the remaining items. Despite this multidimensionality between the MC items and remaining items for Reading and Writing, the effect sizes were weak and did not warrant further investigation.

6.6 ITEM RESPONSE THEORY ANALYSES

Chapter 5, subsection 5.1, introduced IRT and gave a thorough description of the topic. It was noted there that all 2006-07 NECAP items were calibrated using IRT and that the calibrated item parameters were ultimately used to scale both the items and students onto a common framework. The results of those analyses are presented in this subsection and Appendix G.

The tables in Appendix G give the IRT item parameters of all common items on the 2006-07 NECAP tests, broken down by grade and content area. Graphs of the corresponding Test Characteristic Curves (TCCs) and Test Information Functions (TIFs), defined below, accompany the data tables.

TCCs display the expected (average) raw score associated with each θ_j value between –4 and 4. Mathematically, the TCC is computed by summing the ICCs of all items that contribute to the raw score. Using the notation introduced in subsection 5.1, the expected raw score at a given value of θ_j is

$$E(X \mid \theta_j) = \sum_{i=1}^{n} P_i\left(1 \middle| \theta_j\right),$$

where i indexes the items (and n is the number of items contributing to the raw score),

j indexes students (here, θ_j runs from –4 to 4)
$E(X \mid \theta_j)$ is the expected raw score for a student of ability θ_j.

The expected raw score monotonically increases with θ_j, consistent with the notion that students of high ability tend to earn higher raw scores than do students of low ability. Most TCCs are "S-shaped" – flatter at the ends of the distribution and steeper in the middle.

The TIF displays the amount of statistical information that the test provides at each value of θ_j. There is a direct relation between the information of a test and its standard error of measurement (SEM). Information functions depict test precision across the entire latent trait continuum. For long tests, the SEM at a given θ_j is approximately equal to the inverse of the square root of the statistical information at θ_j (Hambleton, Swaminathan, & Rogers, 1991):

$$SEM(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}}$$

Compared to the tails, TIFs are often higher near the middle of the θ distribution, where most students are located and most items are sensitive by design.

6.7 EQUATING RESULTS

As discussed in Section 5.1, a combination of IRT models was used for scaling NECAP items: 3PL for dichotomously scored items; 3PL with c=0 (i.e., 2PL) for short answer items; GRM for polytomously scored items. As a result of conducting the IRT calibration and the equating process (see Section 5.2), an Equating Report was generated. The Equating Report is included as Appendix C of this technical report.

There were three basic steps involved in the equating and scaling activities: IRT calibrations, identification of equating items, and execution of the Stocking & Lord equating procedure. These, along with the various quality control procedures implemented within the Psychometrics Department at Measured Progress, have been reviewed with the NECAP state testing directors and the NECAP Technical Advisory Committee.

IRT CALIBRATION RESULTS

All IRT calibrations were conducted using the PARSCALE (v 4.1) software system. Details of the calibration process are included in Section I.c of the Equating Report (Appendix C). As seen in Table I.c.1 of the Equating Report, all IRT calibrations converged within 100 Newton cycles. Table I.c.2 summarizes required interventions during the calibration process. Interventions consisted of either fixing the lower asymptote (i.e., the c-parameter) or changing the initial estimate for the c-parameter to some other value. For example in grade 3 reading, four items required intervention. Two items required fixing the c-parameters, two required different initial values for c-parameters. Similar numbers of items required intervention across the various grade-contents of NECAP.

Section I.d of the Equating Report outlines several other quality control activities undertaken during the IRT calibrations. For example, all items were evaluated to ensure that the a-parameters (i.e., item discrimination parameters) were not unreasonably low, and that the standard errors (SE) on the

estimated b-parameters (i.e., the item difficulty parameters) were not too high. Low a-parameters and/or high SE b-parameter values sometimes indicate multidimensionality or other violations of IRT models. In this particular analysis, no items across all grade-content combinations were problematic suggesting that reasonable model-data fit was obtained for the NECAP program.

A considerable amount of time at Measured Progress was spent on evaluating item-level fit. For each item, observed results conditional on the performance continuum (θ) were plotted against model-estimated item parameters. Differences between observed and modeled values were used to evaluate model fit subjectively. This method helped determine starting c-parameter values (Table I.c.2) and ensured that all item parameter estimates resulted in item characteristic curves that accurately represented the student-test question interaction.

IDENTIFICATION OF EQUATING ITEMS

Through the test development process and specifications in test blueprint documents, psychometricians at Measured Progress located the specific equating items used in the NECAP program. These equating items serve to link this year's psychometric scale (i.e., the θ scale) to the previous year's scale. The delta analysis procedure was then used to evaluate the equating items. For the delta procedure, the p-values of the equating items from the current and the previous year's administrations were transformed to the commonly used ETS delta metric. A scatter plot of the delta values from the two administrations was formed and a trend line established. Items whose perpendicular distance to the trend line was more than three standard deviation units were not included in calculating the Stocking & Lord transformation constants used in the equating process.

Table I.c.3 of the Equating Report contains a list of all equating items that were removed from the analysis. For example in grade 4 math, item 227082 was not used as part of the equating solution since it was 3.165 standard deviations from the trend line. Table I.c.3 lists other items evaluated during the equating process and specifies any actions required. Typically, about 3 or 4 items for each grade-content are evaluated during this type of analysis; the results in Table I.c.3 very typical for a program

such as NECAP.

Section II.b of the Equating Report contains the results from the rescore analysis conducted on the polytomously scored equating items. For this analysis, a random set of papers from the previous year's administration were interspersed with this year's papers in order to investigate what effect, if any, scorers were having on the equating items. Both effect size and absolute differences were studied, and results are presented in the Equating Report. As is seen in the grade-content tables, no polytomously scored items were discarded from use as equating items.

STOCKING & LORD RESULTS

Table I.e.1 of the Equating Report presents the Stocking & Lord transformation constants used for each grade-content in the NECAP program. These constants are analogous to slope (labeled "A") and intercept (labeled "B") terms, and are used to place the item parameters estimated in the calibrations discussed above onto the previous year's scale. Ideally, equatings are conducted on parallel test forms, and the adjustment made in the equating process in minimal. From this perspective, the expectation is for the A constant to be 1.0 and the B constant to be 0.0. The NECAP values presented in Table I.e.1 are all within a very reasonable range, and the largest values were found in grade 8 reading (A=1.085364 and B=-0.217158). Though these values were larger than the other values in the table, they are still quite reasonable within the Stocking & Lord equating framework. Nonetheless, psychometricians at Measured Progress did focus carefully on the grade 8 reading results to ensure that proper model fit was established after the equating process. No additional steps were required grade 8 reading, and the resulting transformation constants shown in Table I.e.1 were used in the equating process.

CHAPTER 7—RELIABILITY

Although an individual item's performance is an important focus for evaluation, a complete evaluation of a test must also address the way in which items function together and complement one another. Any measurement includes some amount of measurement error. No academic test can measure student performance with perfect accuracy; some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. Items that function well together produce tests that have less measurement error (i.e., the error is small on average). Such tests are described as "reliable."

There are a number of ways to estimate a test's reliability. One approach is to split all test items into two groups and then correlate students' scores on the two half-tests. This is known as a *split-half* estimate of reliability. If the two half-test scores correlate highly, items on the two half-tests are likely measuring very similar knowledge or skills. Such a correlation is evidence that the items complement one another and suggest that measurement error will be minimal.

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation. Cronbach (1951) provided a statistic, alpha (α), which avoids this concern of the split-half method. By comparing individual item variances to total test variance, Cronbach's α coefficient estimates the average of all possible split-half reliability coefficients and was used to assess the reliability of the 2006-07 NECAP tests:

$$\alpha \equiv \frac{n}{n-1}\left[1-\frac{\sum_{i=1}^{n}\sigma^2\left(Y_i\right)}{\sigma_x^2}\right]$$

where *i* indexes the item,

n is the number of items,
$\sigma^2(Y_i)$ represents individual item variance
σ_x^2 represents the total test variance.

7.1 RELIABILITY AND STANDARD ERRORS OF MEASUREMENT

Table 7-1 presents descriptive statistics, Cronbach's α coefficient, and raw score standard errors of measurement (SEMs) for each content area and grade (statistics are based on common items only).

Table 7-1. 2006-07 NECAP Common Item Raw Score Descriptive Statistics, Reliabilities, and Standard Errors of Measurement by Grade and Subject Area.

Grade	Subject	N	Possible Score	Min Score	Max Score	Mean Score	Score SD	Reliability (α)	S.E.M.
3	Math	32197	65	0	65	41.267	13.035	0.929	3.465
	Reading	32117	52	0	52	35.127	10.601	0.889	3.525
4	Math	32346	65	0	65	39.254	12.833	0.927	3.460
	Reading	32218	52	0	52	33.472	9.6080	0.889	3.203
5	Math	32779	66	0	66	34.200	14.063	0.917	4.055
	Reading	32687	52	0	52	29.916	9.119	0.893	2.988
	Writing	32626	37	0	37	21.134	5.170	0.750	2.585
6	Math	33874	66	0	66	33.198	14.728	0.926	4.010
	Reading	33756	52	0	52	31.798	9.123	0.889	3.042
7	Math	35210	66	0	66	29.441	12.957	0.902	4.051
	Reading	35122	52	0	52	30.802	8.940	0.892	2.940
8	Math	35415	66	0	66	27.727	13.391	0.915	3.893
	Reading	35318	52	0	52	32.412	9.496	0.897	3.054
	Writing	35167	37	0	37	22.965	6.109	0.760	2.993

For mathematics, the reliability coefficient ranged from 0.90 to 0.93, for reading 0.89 to 0.90. For the grade 5 and grade 8 writing tests, the values were 0.75 and 0.76, respectively. Because different grades and content areas have different test designs (e.g., the number of items varies by test), it is inappropriate to make inferences about the quality of one test by comparing its reliability to that of another test from a different grade and/or content area.

7.2 SUBGROUP RELIABILITY

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2006-07 NECAP tests. Appendix H presents reliabilities for various subgroups of interest. Subgroup Cronbach's α's were calculated using the formula defined above using only the members of the subgroup in question in the computations. For mathematics, subgroup reliabilities ranged from 0.86 to 0.94, for reading from 0.84 to 0.92, and for writing from 0.67 to 0.82. The subgroup reliabilities for writing were lower than those for the other two content areas, but the two writing tests (grades 5 and 8) were consistent with each other.

For several reasons, the results of this subsection should be interpreted with caution. First, inherent differences between grades and content areas preclude making valid inferences about the quality of a test based on statistical comparisons with other tests. Second, reliabilities are dependent not only on the measurement properties of a test but on the statistical distribution of the studied subgroup. For example, subgroup sample sizes may vary considerably (see Appendix H), which results in natural variation in reliability coefficients. Or α, which is a type of correlation coefficient, may be artificially depressed for subgroups with little variability (Draper & Smith, 1998). Third, there is no industry standard to interpret the strength of a reliability coefficient, and this is particularly true when the population of interest is a single subgroup.

7.3 STRATIFIED COEFFICIENT ALPHA

According to Feldt and Brennan (1989), a prescribed distribution of items over categories (such as different item types) indicates the presumption that at least a small, but important, degree of unique variance is associated with the categories. In contrast, Cronbach's α coefficient is built on the assumption that there are no such local or clustered dependencies. A stratified version of coefficient α corrects for this problem.

The formula for stratified α is as follows:

$$\alpha_{strat} = 1 - \frac{\sum_{j=1}^{k} \sigma_{x_j}^2 (1-\alpha)}{\sigma_x^2}$$

where *j* indexes the subtests or categories,

$\sigma_{x_j}^2$ represents the variance of the *k* individual subtests or categories,
α is the unstratified Cronbach's α coefficient, and
σ_x^2 represents the total test variance.

Stratified α was calculated separately for each grade/content combination. The results of stratification based on item type (MC versus OR) are presented below in Table 7-2. This is directly followed by results of stratification based on form in Table 7-3.

Table 7-2. 2006-07 NECAP: Common Item α and Stratified α by Grade, Subject, and Item Type.

		All	MC		OR		
Grade	**Subject**	α	α	**N**	α	**N (poss)**	**Stratified α**
3	Math	0.93	0.89	35	0.85	20 (30)	0.93
	Reading	0.89	0.89	28	0.75	6 (24)	0.91
4	Math	0.93	0.88	35	0.85	20 (30)	0.93
	Reading	0.89	0.87	28	0.76	6 (24)	0.90
5	Math	0.92	0.88	32	0.84	16 (34)	0.92
	Reading	0.89	0.87	28	0.84	6 (24)	0.91
6	Math	0.93	0.86	32	0.88	16 (34)	0.93
	Reading	0.89	0.85	28	0.84	6 (24)	0.91
7	Math	0.9	0.83	32	0.83	16 (34)	0.91
	Reading	0.89	0.85	28	0.87	6 (24)	0.91
8	Math	0.92	0.84	32	0.86	16 (34)	0.92
	Reading	0.9	0.86	28	0.88	6 (24)	0.92

All = MC and OR; MC = multiple-choice; OR = open response
N = number of items; poss = total possible open-response points

Table 7-3. 2006-07 NECAP: Reliability by Grade, Subject, Item Type, and Form.

Grade	Subject	Stat	Form1	Form2	Form3	Form4	Form5	Form6	Form7	Form8	Form9
3	Math	All α	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
		MC α	0.90	0.91	0.90	0.91	0.91	0.90	0.91	0.90	0.90
		OR α	0.86	0.87	0.88	0.88	0.88	0.87	0.86	0.87	0.88
		Frmt Strat	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
		Com alpha	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.92	0.93
	Reading	All α	0.92	0.92	0.93	0.89	0.89	0.89	0.89	0.89	0.89
		MC α	0.91	0.92	0.93	0.89	0.89	0.88	0.89	0.89	0.89
		OR α	0.82	0.82	0.82	0.75	0.76	0.74	0.74	0.74	0.74
		Frmt Strat	0.93	0.93	0.94	0.90	0.91	0.90	0.90	0.90	0.90
		Com alpha	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
4	Math	All α	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
		MC α	0.90	0.89	0.90	0.89	0.89	0.90	0.90	0.89	0.90
		OR α	0.88	0.88	0.87	0.87	0.88	0.88	0.87	0.88	0.87
		Frmt Strat	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
		Com alpha	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	Reading	All α	0.92	0.92	0.92	0.89	0.89	0.89	0.89	0.89	0.89
		MC α	0.91	0.91	0.90	0.87	0.87	0.87	0.87	0.87	0.87
		OR α	0.83	0.82	0.81	0.76	0.76	0.77	0.76	0.75	0.77
		Frmt Strat	0.93	0.93	0.93	0.90	0.90	0.90	0.90	0.90	0.90
		Com alpha	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
5	Math	All α	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
		MC α	0.88	0.89	0.89	0.90	0.89	0.89	0.89	0.89	0.89
		OR α	0.88	0.88	0.87	0.88	0.87	0.88	0.88	0.88	0.87
		Frmt Strat	0.93	0.94	0.93	0.94	0.93	0.94	0.94	0.94	0.94
		Com alpha	0.92	0.92	0.92	0.92	0.91	0.92	0.92	0.92	0.92
	Reading	All α	0.93	0.92	0.93	0.89	0.88	0.89	0.89	0.89	0.90
		MC α	0.91	0.90	0.90	0.87	0.86	0.87	0.87	0.87	0.87
		OR α	0.90	0.89	0.89	0.83	0.82	0.83	0.83	0.84	0.83
		Frmt Strat	0.94	0.94	0.94	0.91	0.90	0.91	0.91	0.91	0.91
		Com alpha	0.89	0.89	0.90	0.89	0.88	0.89	0.89	0.89	0.90
	Writing[1]	All α	0.75								
		MC α	0.70								
		OR α	0.65								
		Frmt Strat	0.76								
		Com alpha	0.75								

cont'd

Table 7-3. 2006-07 NECAP: Reliability by Grade, Subject, Item Type, and Form.

Grade	Subject	Stat	Form1	Form2	Form3	Form4	Form5	Form6	Form7	Form8	Form9
6	Math	All α	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
		MC α	0.88	0.88	0.89	0.88	0.88	0.88	0.89	0.88	0.89
		OR α	0.91	0.90	0.90	0.90	0.91	0.90	0.91	0.91	0.90
		Frmt Strat	0.95	0.94	0.95	0.94	0.94	0.94	0.95	0.94	0.94
		Com alpha	0.93	0.92	0.93	0.93	0.93	0.93	0.93	0.92	0.92
	Reading	All α	0.93	0.92	0.93	0.89	0.88	0.89	0.89	0.89	0.89
		MC α	0.90	0.90	0.90	0.85	0.84	0.85	0.85	0.85	0.85
		OR α	0.89	0.89	0.89	0.85	0.84	0.85	0.85	0.84	0.84
		Frmt Strat	0.94	0.94	0.94	0.91	0.91	0.91	0.91	0.91	0.91
		Com alpha	0.89	0.89	0.89	0.89	0.88	0.89	0.89	0.89	0.89
7	Math	All α	0.92	0.92	0.92	0.93	0.92	0.92	0.92	0.92	0.92
		MC α	0.86	0.86	0.84	0.86	0.86	0.86	0.85	0.86	0.85
		OR α	0.87	0.87	0.87	0.88	0.88	0.87	0.86	0.87	0.88
		Frmt Strat	0.92	0.93	0.92	0.93	0.93	0.92	0.92	0.93	0.93
		Com alpha	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
	Reading	All α	0.93	0.93	0.93	0.89	0.89	0.89	0.89	0.89	0.89
		MC α	0.90	0.90	0.90	0.85	0.85	0.85	0.85	0.85	0.85
		OR α	0.91	0.91	0.91	0.87	0.87	0.86	0.87	0.87	0.86
		Frmt Strat	0.94	0.94	0.94	0.91	0.91	0.91	0.91	0.91	0.91
		Com alpha	0.90	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
8	Math	All α	0.94	0.93	0.93	0.93	0.93	0.94	0.94	0.93	0.93
		MC α	0.87	0.87	0.87	0.86	0.86	0.88	0.87	0.87	0.87
		OR α	0.90	0.89	0.89	0.89	0.89	0.90	0.90	0.89	0.89
		Frmt Strat	0.94	0.94	0.94	0.93	0.93	0.94	0.94	0.94	0.94
		Com alpha	0.92	0.91	0.92	0.92	0.91	0.92	0.92	0.91	0.91
	Reading	All α	0.93	0.93	0.93	0.90	0.89	0.90	0.89	0.90	0.90
		MC α	0.90	0.90	0.91	0.86	0.85	0.87	0.85	0.86	0.85
		OR α	0.92	0.91	0.92	0.88	0.88	0.88	0.87	0.88	0.87
		Frmt Strat	0.95	0.94	0.95	0.92	0.92	0.93	0.92	0.92	0.92
		Com alpha	0.90	0.90	0.90	0.90	0.89	0.90	0.89	0.90	0.90
	Writing[1]	All α	0.76								
		MC α	0.70								
		OR α	0.69								
		Frmt Strat	0.78								
		Com alpha	0.76								

MC = multiple-choice; OR = open response; All = MC and OR
All α = common and matrix items; MC α = MC items only; OR α = OR items only; Frmt Strat = stratified by MC/OR; Com alpha = common items only
[1]Writing tests had only one form

Not surprisingly, reliabilities were higher on the full test than on subsets of items (i.e., only MC or OR items).

7.4 REPORTING SUBCATEGORIES RELIABILITY

In subsection 7.3, the reliability coefficients were calculated based on form and item type. Item type represents just one way of breaking an overall test into subtests. Of even more interest are reliabilities for the reporting subcategories within NECAP subject areas, described in Chapter 2. Cronbach's α coefficients for subcategories were calculated via the same formula defined in subsection 7.1 using just the items of a given subcategory in the computations. Results are presented in Table 7-4. Once again as expected, because they are based on a subset of items rather than the full test, computed subcategory reliabilities were lower (sometimes substantially so) than were overall test reliabilities, and interpretations should take this into account.

Table 7-4. 2006-07 NECAP Common Item α by Grade, Subject, and Reporting Subcategory.

Grade	Subject	Reporting Subcategory	Possible Points	α
3	Math	Number & Operations	35	0.89
		Geometry & Measurement	10	0.61
		Functions & Algebra	10	0.69
		Data, Statistics, & Probability	10	0.70
	Reading	Word ID/Vocabulary	19	0.73
		Literary	16	0.70
		Informational	17	0.74
		Initial Understanding	20	0.77
		Analysis & Interpretation	13	0.64
4	Math	Number & Operations	32	0.87
		Geometry & Measurement	13	0.67
		Functions & Algebra	10	0.66
		Data, Statistics, & Probability	10	0.70
	Reading	Word ID/Vocabulary	20	0.76
		Literary	14	0.65
		Informational	18	0.76
		Initial Understanding	20	0.78
		Analysis & Interpretation	12	0.60
5	Math	Number & Operations	30	0.85
		Geometry & Measurement	14	0.63
		Functions & Algebra	12	0.63
		Data, Statistics, & Probability	10	0.63
	Reading	Word ID/Vocabulary	10	0.69
		Literary	22	0.78
		Informational	20	0.76
		Initial Understanding	22	0.78
		Analysis & Interpretation	20	0.74
	Writing	Structures of Language & Writing Conventions	10	0.70
		Short Responses	12	0.70
		Extended Responses	15	0.17
6	Math	Number & Operations	26	0.86
		Geometry & Measurement	17	0.70
		Functions & Algebra	13	0.67
		Data, Statistics, & Probability	10	0.65
	Reading	Word ID/Vocabulary	10	0.68
		Literary	21	0.77
		Informational	21	0.75
		Initial Understanding	21	0.77
		Analysis & Interpretation	21	0.74

cont'd

Table 7-4. 2006-07 NECAP Common Item α by Grade, Subject, and Reporting Subcategory.

Grade	Subject	Reporting Subcategory	Possible Points	α
7	Math	Number & Operations	20	0.77
		Geometry & Measurement	16	0.59
		Functions & Algebra	19	0.73
		Data, Statistics, & Probability	11	0.63
	Reading	Word ID/Vocabulary	10	0.65
		Literary	22	0.79
		Informational	20	0.77
		Initial Understanding	18	0.75
		Analysis & Interpretation	24	0.78
8	Math	Number & Operations	13	0.69
		Geometry & Measurement	16	0.69
		Functions & Algebra	27	0.83
		Data, Statistics, & Probability	10	0.63
	Reading	Word ID/Vocabulary	10	0.70
		Literary	20	0.77
		Informational	22	0.78
		Initial Understanding	17	0.73
		Analysis & Interpretation	25	0.80
	Writing	Structures of Language & Writing Conventions	10	0.70
		Short Responses	12	0.78
		Extended Responses	15	0.18

For mathematics, subcategory reliabilities ranged from 0.59 to 0.89, for reading from 0.60 to 0.79, and for writing from 0.17 to 0.78. The subcategory reliabilities for the Extended Response writing categories were lower than those of other categories because 12 of the 15 points for the category came from a single 12-point writing prompt item. In general, the subcategory reliabilities were lower than those based on the total test and approximately to the degree one would expect based on classical test theory. Qualitative differences between grades and content areas once again preclude valid inferences about the quality of the full test based on statistical comparisons among subtests.

7.5 RELIABILITY OF ACHIEVEMENT LEVEL CATEGORIZATION

All test scores contain measurement error; thus, classifications based on test scores are also subject to measurement error. After the 2006-07 NECAP achievement levels were specified and students classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications. For every 2006-07 NECAP grade and content area, each student was

classified into one of the following achievement levels: Substantially Below Proficient (SBP), Partially Proficient (PP), Proficient (P), or Proficient With Distinction (PWD). This section of the report explains the methodologies used to assess the reliability of classification decisions and presents the results.

ACCURACY AND CONSISTENCY

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated, because errorless test scores do not exist.

Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are given to the same group of students. In operational test programs, however, such a design is usually impractical. Instead, techniques, such as one due to Livingston and Lewis (1995), have been developed to estimate both the accuracy and consistency of classification decisions based on a single administration of a test. The Livingston and Lewis technique was used for the 2006-07 NECAP because it is easily adaptable to tests of all kinds of formats, including mixed-format tests.

CALCULATING ACCURACY

The accuracy and consistency estimates reported below make use of "true scores" in the classical test theory sense. A true score is the score that would be obtained if a test had no measurement error. Of course, true scores cannot be observed and so must be estimated. In the Livingston and Lewis method, estimated true scores are used to classify students into their "true" achievement level.

For the 2006-07 NECAP, after various technical adjustments were made (described in Livingston and Lewis, 1995), a 4 x 4 contingency table of accuracy was created for each content area and grade, where cell $[i,j]$ represented the estimated proportion of students whose true score fell into achievement level i (where $i = 1 – 4$) and observed score into achievement level j (where $j = 1 – 4$). The sum of the diagonal entries, i.e., the proportion of students whose true and observed achievement levels

matched one another, signified overall accuracy.

CALCULATING CONSISTENCY

To estimate consistency, true scores were used to estimate the joint distribution of classifications on two independent, parallel test forms. Following statistical adjustments (per Livingston and Lewis, 1995), a new 4 × 4 contingency table was created for each content area and grade and populated by the proportion of students who would be classified into each combination of achievement levels according to the two (hypothetical) parallel test forms. Cell [*i*,*j*] of this table represented the estimated proportion of students whose observed score on the first form would fall into achievement level *i* (where *i* = 1 – 4), and whose observed score on the second form would fall into achievement level *j*(where *j* = 1 – 4). The sum of the diagonal entries, i.e., the proportion of students classified by the two forms into exactly the same achievement level, signified overall consistency.

CALCULATING KAPPA

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

$$\kappa = \frac{\text{(Observed agreement) - (Chance agreement)}}{\text{1 - (Chance agreement)}} = \frac{\sum_i C_{ii} - \sum_i C_{i.}C_{.i}}{1 - \sum_i C_{i.}C_{.i}},$$

where:

$C_{i.}$ is the proportion of students whose observed achievement level would be *Level i* (where *i*=1 – 4) on the first hypothetical parallel form of the test;
$C_{.i}$ is the proportion of students whose observed achievement level would be *Level i* (where *i*=1 – 4) on the second hypothetical parallel form of the test;
C_{ii} is the proportion of students whose observed achievement level would be *Level i* (where *i*=1 – 4) on both hypothetical parallel forms of the test.

Because κ is corrected for chance, its values are lower than are other consistency estimates.

RESULTS OF ACCURACY, CONSISTENCY, AND KAPPA ANALYSES

The accuracy and consistency analyses described above are tabulated in Appendix I. The

appendix includes the accuracy and consistency contingency tables described above and the overall accuracy and consistency indices, including kappa.

Accuracy and consistency values conditional upon achievement level are also given in Appendix I. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, the conditional accuracy value is 0.732 for the PP achievement level for mathematics grade 3. This figure indicates that among the students whose true scores placed them in the PP achievement level, 73.2% of them would be expected to be in the PP achievement level when categorized according to their observed score. Similarly, the corresponding consistency value of 0.642 indicates that 64.2% of students with observed scores in PP would be expected to score in the PP achievement level again if a second, parallel test form were used.

For some testing situations, the greatest concern may be decisions around level thresholds. For example, if a college gave credit to students who achieved an Advanced Placement test score of 4 or 5, but not to scores of 1, 2, or 3, one might be interested in the accuracy of the dichotomous decision below-4 versus 4-or-above. For the 2006-07 NECAP, Appendix I provides accuracy and consistency estimates at each cutpoint as well as false positive and false negative decision rates. (False positives are the proportion of students whose observed scores were above the cut and true scores below the cut. False negatives are the proportion of students whose observed scores were below the cut and true scores above the cut.)

The above indices are derived from Livingston & Lewis' (1995) method of estimating the accuracy and consistency of classifications. It should be noted that Livingston & Lewis discuss two versions of the accuracy and consistency tables. A standard version performs calculations for forms parallel to the form taken. An "adjusted" version adjusts the results of one form to match the observed score distribution obtained in the data. The tables reported in Appendix I use the standard version for two reasons: 1) this "unadjusted" version can be considered a smoothing of the data, thereby decreasing the variability of the results; and 2) for results dealing with the consistency of two parallel forms, the

unadjusted tables are symmetric, indicating that the two parallel forms have the same statistical properties. This second reason is consistent with the notion of forms that are parallel, i.e., it is more intuitive and interpretable for two parallel forms to have the same statistical distribution as one another.

Descriptive statistics relating to the decision accuracy and consistency of the 2006-07 NECAP tests can be derived from Appendix I. For mathematics, overall accuracy ranged from 0.778 to 0.815; overall consistency ranged from 0.701 to 0.743; the kappa statistic ranged from 0.577 to 0.631. For reading, overall accuracy ranged from 0.781 to 0.818; overall consistency ranged from 0.704 to 0.747; the kappa statistic ranged from 0.542 to 0.622. Finally, for writing, overall accuracy was 0.617 or 0.642 in the two grades tested; overall consistency was 0.516 or 0.539; the kappa statistic was 0.343 or 0.362.

Table 7-5 below summarizes most of the results of Appendix I at a glance. As with other types of reliability, it is inappropriate when analyzing the decision accuracy and consistency of a given test to compare results between grades and content areas.

Table 7-5. 2006-07 NECAP Summary of Decision Accuracy (and Consistency) Results.

		Conditional on Level				At Cut Point		
Content/Grade	**Overall**	**SBP**	**PP**	**P**	**PWD**	**SBP:PP**	**PP:P**	**P:PWD**
Math/3	.81(.74)	.85(.78)	.73(.64)	.82(.77)	.87(.77)	.96(.94)	.93(.90)	.93(.90)
Math/4	.81(.74)	.86(.80)	.72(.62)	.84(.79)	.84(.70)	.95(.94)	.93(.90)	.93(.91)
Math/5	.80(.73)	.82(.75)	.63(.52)	.84(.79)	.87(.76)	.94(.92)	.92(.89)	.94(.92)
Math/6	.81(.74)	.84(.78)	.63(.52)	.84(.79)	.88(.79)	.94(.92)	.93(.90)	.94(.92)
Math/7	.78(.70)	.82(.75)	.56(.45)	.82(.76)	.88(.77)	.92(.89)	.91(.87)	.94(.92)
Math/8	.79(.72)	.83(.77)	.57(.46)	.84(.78)	.88(.77)	.92(.88)	.91(.88)	.95(.94)
Reading/3	.78(.70)	.83(.76)	.69(.59)	.81(.77)	.77(.61)	.96(.95)	.93(.90)	.89(.85)
Reading/4	.79(.71)	.80(.70)	.72(.63)	.79(.74)	.87(.74)	.96(.94)	.91(.87)	.92(.89)
Reading/5	.80(.73)	.80(.70)	.73(.65)	.82(.76)	.87(.76)	.96(.94)	.91(.88)	.93(.90)
Reading/6	.81(.74)	.80(.71)	.73(.64)	.84(.79)	.86(.74)	.96(.94)	.91(.88)	.94(.91)
Reading/7	.82(.75)	.80(.70)	.76(.69)	.84(.80)	.86(.71)	.96(.94)	.91(.87)	.95(.92)
Reading/8	.82(.75)	.82(.74)	.76(.68)	.84(.80)	.86(.73)	.96(.94)	.92(.88)	.95(.93)
Writing/5	.62(.52)	.74(.62)	.49(.41)	.60(.50)	.79(.58)	.88(.83)	.83(.77)	.89(.85)
Writing/8	.64(.54)	.74(.63)	.57(.49)	.62(.53)	.80(.54)	.89(.84)	.84(.78)	.89(.85)

SBP = Substantially Below Proficient; PP = Partially Proficient; P = Proficient; PWD = Proficient with Distinction

CHAPTER 8—VALIDITY

Because interpretations of test scores, and not a test itself, are evaluated for validity, the purpose of the 2006-07 NECAP Technical Report is to describe several technical aspects of the NECAP tests in support of score interpretations (AERA, 1999). Each chapter contributes an important component in the investigation of score validation: test development and design; test administration; scoring, scaling, and equating; item analyses; reliability; and score reporting.

The NECAP tests are based on and aligned with the content standards and performance indicators in the GLEs for mathematics, reading, and writing. Inferences about student achievement on the content standards are intended from NECAP results, which in turn serve evaluation of school accountability and inform the improvement of programs and instruction.

The *Standards for Educational and Psychological Testing* (1999) provides a framework for describing sources of evidence that should be considered when evaluating validity. These sources include evidence on the following five general areas: test content, response processes, internal structure, consequences of testing, and relationship to other variables. Although each of these sources may speak to a different *aspect* of validity, they are not distinct *types* of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

A measure of test content validity is to determine how well the test tasks represent the curriculum and standards for each subject and grade level. This is informed by the item development process, including how test blueprints and test items align with the curriculum and standards. Validation through the content lens was extensively described in Chapter 2. Item alignment with content standards; item bias; sensitivity and content appropriateness review processes; adherence to the test blueprint; use of multiple item types; use of standardized administration procedures, with accommodated options for participation; and appropriate test administration training are all components of validity evidence based on test content.

All NECAP test questions were aligned by educators with specific content standards and underwent several rounds of review for content fidelity and appropriateness. Items were presented to students in multiple formats (MC, SA, and CR). Finally, tests were administered according to mandated standardized procedures, with allowable accommodations, and all test coordinators and test administrators were required to familiarize themselves with and adhere to all of the procedures outlined in the *NECAP Test Coordinator* and *Test Administrator* manuals.

The scoring information in Chapter 4 described both the steps taken to train and monitor hand-scorers and quality control procedures related to scanning and machine-scoring. Additional studies might be helpful for evidence on student response processes. For example, think-aloud protocols could be used to investigate students' cognitive processes when confronting test items.

Evidence on internal structure was extensively detailed in discussions of scaling and equating, item analyses, and reliability in Chapters 5, 6, and 7. Technical characteristics of the internal structure of the tests were presented in terms of classical item statistics (item difficulty and item-test correlation), differential item functioning analyses, a variety of reliability coefficients, SEM, multidimensionality hypothesis testing and effect size estimation, and IRT parameters and procedures. In general, item difficulty indices were within acceptable and expected ranges; very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicated that students who performed well on individual items tended to perform well overall. Chapter 5 also described the method used to equate the 2006-07 test to the 2005-06 scales.

Evidence on the consequences of testing was addressed in information on scaled score and reporting in Chapters 5 and 9 and in the *Guide to Using the 2006 NECAP Reports*, which is a separate document referenced in the discussion of reporting. Each of these spoke to efforts undertaken for providing the public with accurate and clear test score information. Scaled scores simplify results reporting across content areas, grade levels, and successive years. Achievement levels give reference points for mastery at each grade level, another useful and simple way to interpret scores. Several

different standard reports were provided to stakeholders. Evidence on the consequences of testing could be supplemented with broader research on the impact on student learning of NECAP testing.

8.1 QUESTIONNAIRE DATA

A measure of external validity was provided by comparing student performance with answers to a questionnaire administered at the end of test. The questionnaire contained 29 questions: Nine concerned the content area of reading, ten concerned mathematics, and ten concerned writing. Most of the questions were designed to gather information about students and their study habits; however, a subset could be utilized in the test of external validity. One question was chosen from each content area that was most expected to correlate with student performance on NECAP tests. To the extent that the answers to those questions did correlate with student performance in the anticipated manner, the external validity of score interpretations was confirmed. The three questions are now discussed one at a time.

Question 8, concerning reading, read as follows:

8. How often do you choose to read in your free time?
   A. almost every day
   B. a few times a week
   C. a few times a month
   D. I almost never read.

It was anticipated that students who read more in their free time would have higher average scaled scores and achievement level designations in reading than students who did not read as much. In particular, it was expected that on average, reading performance among students who chose "A" would meet or exceed performance of students who chose "B," whose performance would meet or exceed that of students who chose "C," whose performance would meet or exceed that of students who chose "D." This pattern was observed in Table 8-1 in all grades, both in terms of average scaled scores and the percentage of students in the *Proficient with Distinction* achievement level.

Table 8-1. 2006-07 NECAP: Average Scaled Score, and Counts and Percentages within Performance Levels, of Responses to Item 8[1] of Student Questionnaire – Reading.

Grade	Resp	Number Resp	Percentage Resp	Avg SS	N SBP	N PP	N P	N PWD	% SBP	% PP	% P	% PWD
	(blank)	3752	12	343	674	745	1829	504	18	20	49	13
	A	16245	51	347	1508	2576	9115	3046	9	16	56	19
3	B	7804	24	346	826	1270	4534	1174	11	16	58	15
	C	1773	6	344	289	332	931	221	16	19	53	12
	D	2544	8	340	545	586	1240	173	21	23	49	7
	(blank)	3200	10	442	561	711	1488	440	18	22	47	14
	A	15210	47	446	1353	2718	8007	3132	9	18	53	21
4	B	9451	29	445	873	1999	5241	1338	9	21	55	14
	C	1936	6	442	298	461	992	185	15	24	51	10
	D	2421	8	438	531	683	1095	112	22	28	45	5
	(blank)	2947	9	542	530	596	1372	449	18	20	47	15
	A	14433	44	547	1222	2231	7884	3096	8	15	55	21
5	B	10355	32	544	1036	2128	5796	1395	10	21	56	13
	C	2355	7	542	363	562	1190	240	15	24	51	10
	D	2597	8	538	590	787	1110	110	23	30	43	4
	(blank)	3587	11	641	710	864	1689	324	20	24	47	9
	A	11585	34	649	790	1704	6793	2298	7	15	59	20
6	B	11624	34	645	1056	2509	6777	1282	9	22	58	11
	C	3543	10	643	436	864	1990	253	12	24	56	7
	D	3417	10	639	693	998	1611	115	20	29	47	3
	(blank)	4098	12	741	794	1143	1853	308	19	28	45	8
	A	9335	27	749	538	1489	5568	1740	6	16	60	19
7	B	11684	33	745	932	3037	6606	1109	8	26	57	9
	C	4723	13	742	541	1403	2547	232	11	30	54	5
	D	5282	15	739	832	1949	2364	137	16	37	45	3
	(blank)	3237	9	839	832	818	1305	282	26	25	40	9
	A	8837	25	849	576	1329	5222	1710	7	15	59	19
8	B	11032	31	845	1044	2502	6335	1151	9	23	57	10
	C	5472	15	842	663	1607	2874	328	12	29	53	6
	D	6740	19	838	1239	2381	2945	175	18	35	44	3

[1]Question 8: How often do you choose to read in your free time? A = almost every day; B = a few times a week; C = a few times a month; D = I almost never read.
SBP = Substantially Below Proficient; PP = Partially Proficient; P = Proficient; PWD = Proficient with Distinction.

Question 15, concerning mathematics, read as follows:

15. How often do you have mathematics homework?
    A. almost every day
    B. a few times a week
    C. a few times a month
    D. I usually don't have homework in mathematics.

As anticipated, the relationship between Question 15 and student performance in mathematics (see Table 8-2 below) mirrored the pattern of Question 8 at each grade: On average, mathematics performance among students who chose "A" met or exceeded the performance of students who chose "B," whose performance met or exceeded that of students who chose "C," whose performance met or exceeded that of students who chose "D." This pattern was again evident both in terms of average scaled scores and the percentage of students in the *Proficient with Distinction* achievement level.

Table 8-2. 2006-07 NECAP: Average Scaled Score, and Counts and Percentages within Performance Levels, of Responses to Item 15[1] of Student Questionnaire – Mathematics.

Grade	Resp	N Resp	% Resp	Avg SS	N SBP	N PP	N P	N PWD	% SBP	% PP	% P	% PWD
	(blank)	3823	12	341	791	839	1527	666	21	22	40	17
	A	13787	43	344	1772	2846	6357	2812	13	21	46	20
3	B	10338	32	344	1168	2239	4842	2089	11	22	47	20
	C	2104	7	343	317	427	950	410	15	20	45	19
	D	2146	7	340	456	573	863	254	21	27	40	12
	(blank)	3211	10	440	737	731	1348	395	23	23	42	12
	A	15793	49	443	2359	3275	7773	2386	15	21	49	15
4	B	10103	31	443	1519	2288	5010	1286	15	23	50	13
	C	1778	5	442	320	418	825	215	18	24	46	12
	D	1461	5	438	405	369	580	107	28	25	40	7
	(blank)	2942	9	540	782	572	1186	402	27	19	40	14
	A	17752	54	544	2607	3216	8579	3350	15	18	48	19
5	B	9359	29	543	1649	1891	4441	1378	18	20	47	15
	C	1573	5	541	332	306	728	207	21	19	46	13
	D	1153	4	536	391	267	412	83	34	23	36	7
	(blank)	3634	11	638	1110	661	1383	480	31	18	38	13
	A	18862	56	644	3035	3156	8737	3934	16	17	46	21
6	B	9484	28	642	1947	1845	4172	1520	21	19	44	16
	C	1085	3	639	294	218	433	140	27	20	40	13
	D	809	2	634	371	142	238	58	46	18	29	7
	(blank)	4071	12	738	1350	751	1409	561	33	18	35	14
	A	20490	58	743	3654	3726	9277	3833	18	18	45	19
7	B	8959	25	740	2202	1847	3767	1143	25	21	42	13
	C	852	2	736	332	161	278	81	39	19	33	10
	D	838	2	731	445	157	197	39	53	19	24	5
	(blank)	3297	9	835	1370	596	999	332	42	18	30	10
	A	21908	62	842	4500	3896	9780	3732	21	18	45	17
8	B	8234	23	838	2575	1812	3229	618	31	22	39	8
	C	959	3	834	428	219	253	59	45	23	26	6
	D	1017	3	831	578	177	220	42	57	17	22	4

[1]Question 15: How often do you have mathematics homework? A = almost every day; B = a few times a week; C = a few times a month; D = I usually don't have homework in mathematics.
SBP = Substantially Below Proficient; PP = Partially Proficient; P = Proficient; PWD = Proficient with Distinction.

Finally, Question 29, concerning writing, read as follows:

29. What kinds of writing do you do most in school?
    A. I mostly write stories.
    B. I mostly write reports.
    C. I mostly write about things I've read.
    D. I do all kinds of writing.

For Question 29, the only anticipated outcome was that students who selected choice "D," i.e., those who ostensibly had experience in many different kinds of writing, would tend to outperform students who selected any other answer choice. The expected outcome was realized in both grades 5 and 8 (see Table 8-3).

Table 8-3. 2006-07 NECAP: Average Scaled Score, and Counts and Percentages within Performance Levels, of Responses to Item 29[1] of Student Questionnaire – Writing.

Grade	Resp	N Resp	% Resp	Avg SS	N SBP	N PP	N P	N PWD	% SBP	% PP	% P	% PWD
5	(blank)	3818	12	537	1130	1055	1196	437	30	28	31	11
	A	5904	18	538	1511	1818	1946	629	26	31	33	11
	B	3110	10	537	854	971	999	286	27	31	32	9
	C	3018	9	538	765	872	1078	303	25	29	36	10
	D	16776	51	543	2875	4413	6705	2783	17	26	40	17
8	(blank)	4231	12	835	1386	1350	1137	358	33	32	27	8
	A	3953	11	834	1278	1591	944	140	32	40	24	4
	B	5766	16	838	1367	2203	1804	392	24	38	31	7
	C	4160	12	837	986	1535	1341	298	24	37	32	7
	D	17057	49	842	2312	5948	6743	2054	14	35	40	12

[1]Question 29: What kinds of writing do you do most in school? A = I mostly write stories; B = I mostly write reports; C = I mostly write about things I've read; D = I do all kinds of writing.
SBP = Substantially Below Proficient; PP = Partially Proficient; P = Proficient; PWD = Proficient with Distinction.

Based on the foregoing analysis, the relationship between questionnaire data and performance on the NECAP was consistent with expectations of the three questions selected for the investigation of external validity. See Appendix J for a copy of the questionnaire and complete data comparing questionnaire items and test performance.

8.2 VALIDITY STUDIES AGENDA

The remaining part of this chapter describes further studies of validity that are being considered for the future. These studies could enhance the investigations of validity that have already been performed. The proposed areas of validity to be examined fall into four categories: *external validity, convergent and discriminant validity, structural validity, and procedural validity*. These will be discussed in turn.

EXTERNAL VALIDITY

For the 2006-07 NECAP score interpretations, external validity was assessed through cross-tabulations of NECAP test scores with teacher judgments and questionnaire data. Future

investigations could involve additional variables with which to correlate NECAP results. For example, data could be collected on the grades of each student who took the NECAP tests. As with the analysis of teacher judgments and questionnaire data, cross-tabulations of NECAP achievement levels and assigned grades could be created. The average NECAP scaled score could also be computed for each possible assigned grade (A, B, C, etc.). Analysis would focus on the relationship between NECAP scores and grades in the appropriate class (i.e., NECAP mathematics would be correlated with student grades in mathematics, not reading). NECAP scores could also be correlated with other appropriate classroom tests in addition to final grades.

Further evidence of external validity might come from correlating NECAP scores with scores on another standardized test, such as the Iowa Test of Basic Skills (ITBS). As with the study of concordance between NECAP scores and grades, this investigation would compare scores in analogous content areas (e.g., NECAP reading and ITBS reading comprehension). All tests taken by each student would be appropriate to the student's grade level.

CONVERGENT AND DISCRIMINANT VALIDITY

The concepts of convergent and discriminant validity were defined by Campbell and Fiske (1959) as specific types of validity that fall under the umbrella of *construct validity*. The notion of convergent validity states that measures or variables that are intended to align with one another should actually be aligned in practice. Discriminant validity, on the other hand, is the idea that measures or variables that are intended to differ from one another should not be too highly correlated. Evidence for validity comes from examining whether the correlations among variables are as expected in direction and magnitude.

Campbell and Fiske (1959) introduced the study of different *traits* and *methods* as the means of assessing convergent and discriminant validity. Traits refer to the constructs that are being measured (e.g., mathematical ability), and methods are the instruments of measuring them (e.g., a mathematics test or grade). To utilize the framework of Campbell and Fiske, it is necessary that more than one trait and

more than one method be examined. Analysis is performed through the multi-trait/multi-method matrix, which gives all possible correlations of the different combinations of traits and methods. Campbell and Fiske defined four properties of the multi-trait/multi-method matrix that serve as evidence of convergent and discriminant validity:

- The correlation among different methods of measuring the same trait should be sufficiently different from zero. For example, scores on a mathematics test and grades in a mathematics class should be positively correlated.
- The correlation among different methods of measuring the same trait should be higher than that of different methods of measuring different traits. For example, scores on a mathematics test and grades in a mathematics class should be more highly correlated than are scores on a mathematics test and grades in a reading class.
- The correlation among different methods of measuring the same trait should be higher than the same method of measuring different traits. For example, scores on a mathematics test and grades in a mathematics class should be more highly correlated than scores on a mathematics test and scores on an analogous reading test.
- The pattern of correlations should be similar across comparisons of different traits and methods. For example, if the correlation between test scores in reading and writing is higher than the correlation between test scores in reading and mathematics, it is expected that the correlation between grades in reading and writing would also be higher than the correlation between grades in reading and mathematics.

For NECAP, convergent and discriminant validity could be examined by constructing a multi-trait/multi-method matrix and analyzing the four pieces of evidence described above. The traits examined would be mathematics, reading, and writing; different methods would include NECAP score and such variables as grades, teacher judgments, and/or scores on another standardized test.

STRUCTURAL VALIDITY

Though the previous types of validity examine the concurrence between different measures of the same content area, structural validity focuses on the relation between strands *within* a content area, thus supporting *content validity*. Standardized tests are carefully designed to ensure that all appropriate strands of a content area are adequately covered in test, and structural validity is the degree to which related elements of a test are correlated in the intended manner. For instance, it is desired that performance on different strands of a content area be positively correlated; however, as these strands are designed to measure distinct components of the content area, it is reasonable to expect that each strand would contribute a unique component to the test. Additionally, it is desired that the correlation between different item types (MC, SA, and CR) of the same content area be positive.

As an example, an analysis of NECAP structural validity would investigate the correlation between performance in Geometry and Measurement and performance in Functions and Algebra. Additionally, the concordance between performance on MC items and OR items would be examined. Such a study would address the consistency of NECAP tests within each grade and content area. In particular, the dimensionality analyses of Chapter 6 could be expanded to include confirmatory analyses addressing these concerns.

PROCEDURAL VALIDITY

As mentioned earlier, the *NECAP Test Coordinator* and *Test Administrator* manuals delineated the procedures to which all NECAP test coordinators and test administrators were required to adhere. A study of procedural validity would provide a comprehensive documentation of the procedures that were followed throughout the NECAP administration. The results of the documentation would then be compared to the manuals, and procedural validity would be confirmed to the extent that the two are in alignment. Evidence of procedural validity is important because it verifies that the actual administration practices are in accord with the intentions of the design.

Possible instances where discrepancies can exist between design and implementation include the following: A teacher may spiral test forms incorrectly within a classroom; cheating may occur among students; answer documents may be scanned incorrectly. These are examples of *administration error*. A study of procedural validity involves capturing any administration errors and presenting them within a cohesive document for review.

All potential tests of validity that have been introduced in this chapter will be discussed as candidates for action by the NECAP Technical Advisory Committee (NECAP TAC) during 2008-2009. With the advice of the NECAP TAC, the states will develop a short-term (e.g., 1-year) and longer term (e.g., 2-year to 5-year) plan for validity studies.

SECTION III
2006-07 NECAP REPORTING

CHAPTER 9—SCORE REPORTING

9.1 TEACHING YEAR VS. TESTING YEAR REPORTING

The data used for the NECAP Reports are the results of the fall 2006 administration of the NECAP test. However, the NECAP tests are based on the GLEs from the prior year. For example, the Grade 7 NECAP test, administered in the fall of seventh grade, is based on the grade 6 GLEs. Many students therefore receive the instruction they need for the fall test at a different school than where they are currently enrolled. The state Departments of Education determined that access to results information would be valuable to both the school where the student was tested and the school where the student received instruction in order to improve curriculum. To achieve this goal, separate Item Analysis, School and District Results, and School and District Summary reports were created for the "testing" school and the "teaching" school. Every student who participated in the NECAP test was represented in "testing" reports, and most students were also represented in "teaching" reports. In some cases, such as a student who recently moved to the state, it is not possible to provide information about a student in "teaching" reports.

9.2 PRIMARY REPORTS

There were four primary reports for the 2006–07 NECAP:

- Student Report
- Item Analysis Report
- School and District Results Report
- School and District Summary Report

With the exception of the Student Report, all reports were available for schools and districts to

view or download on a password-secure website hosted by Measured Progress. Student-level data files were also available for districts to download from the secure Web site. Each of these reports is described in the following subsections. Sample reports are provided in Appendix K.

9.3 STUDENT REPORT

The *NECAP Student Report* is a single-page two-sided report that is divided into three sections. The front side of the report includes a letter from the commissioner of education, a description of the achievement levels, and a graph showing state summary results. The reverse side of the student report provides a complete picture of an individual student's performance on the NECAP, in three sections. The first section gives the student's overall performance for each content area. The student's achievement levels and scaled scores are presented numerically as well as in a graphic that places the student's scaled score, with its standard error of measurement bar constructed about it, within the full range of possible scaled scores demarcated into the four achievement levels.

The second section of the report displays the student's achievement level in each content area relative to the percentage of students at each achievement level across the school, district, and state.

The third section of the report shows the student's performance compared to school, district, and statewide performances. Each content area is reported by subcategories. For **reading**, with the exception of Word ID/Vocabulary items, items are reported by Type of Text (Literary, Informational) and Level of Comprehension (Initial Understanding, Analysis and Interpretation). For **mathematics**, the subcategories are Numbers and Operations; Geometry and Measurement; Functions and Algebra; and Data, Statistics, and Probability. The content area subcategories for **writing** are reported on the Structures of Language and Writing Conventions, displayed in the student's writing and in response to MC items, and by the type of response—short or extended.

Student performances by subject area are reported in the context of possible points; average points earned for the school, district, and state; and the average points earned by students at the Proficient level on the total test.

To provide a more complete picture of the student's performance on the writing test in grades 5 and 8, each scorer chose up to three comments about the student's writing performance from a predetermined list produced by the writing representatives from each state department of education. Scorers' comments are presented next to the writing results.

The *NECAP Student Report* is confidential and should be kept secure within the school and district. The Family Educational Rights and Privacy Act (FERPA) requires that access to individual student results be restricted to the student, the student's parents/guardians, and authorized school personnel.

9.4 ITEM ANALYSIS REPORTS

The *NECAP Item Analysis Report* provides a roster of all the students in each school and their performances on the common items in the test, one report per content area. The student names are listed as row headers down the left side of the report, and items are listed as column headers across the top in the order they appeared the released item documents (not the position in which they appeared on the test). For each item, seven pieces of information are shown: the released item number, the content strand for the item, the GLE code for the item, the Depth of Knowledge code for the item, the item type, the correct response letter for MC items, and the total possible points for each item. For each student, MC items are marked either with a plus sign (+), indicating that the student chose the correct response, or a letter (from A to D), indicating the incorrect response chosen by the student. For CR items, the number of points that the student attained is shown. All responses to released items are shown is the report, regardless of the student's participation status.

The columns on the right side of the report show Total Test Results broken into several categories. The Subcategory Points Earned columns show points earned by the student in each content area relative to total points possible. The Total Points Earned column is a summary of all points earned and total possible points in the content area. The last two columns show the Scaled Score and Achievement Level for each student. For students who are reported as Not Tested, a code appears in the

Achievement Level column to indicate the reason why the student did not test. The descriptions of these codes can be found on the legend, after the last page of data on the report. It is important to note that not all items used to compute student scores are included in this report. Only those items that have been released are included. At the bottom of the report, the average percentage correct for each MC item and average scores for the SA and CR items and writing prompts is shown across the school, district, and state.

The *NECAP Item Analysis Report* is confidential and should be kept secure within the school and district. The FERPA requires that access to individual student results be restricted to the student, the student's parents/guardians, and authorized school personnel.

9.5 SCHOOL AND DISTRICT RESULTS REPORTS

The *NECAP School Results Report* and the *NECAP District Results Report* consist of three parts: the grade level summary report (page 2), the content area results (pages 3, 5, and 7), and the disaggregated content area results (pages 4, 6, and 8).

The grade level summary report provides a summary of participation in the NECAP and a summary of NECAP results. The participation section on the top half of the page shows the number and percentage of students who were enrolled as of October 1, 2006-07. The total number of students enrolled is defined as the number of students tested plus the number of students not tested.

Because students who were not tested did not participate, average school scores were not affected by non-tested students. These students were included in the calculation of the percentage of students participating but not in the calculation of scores. For students who participated in some but not all sessions of the NECAP test, actual scores were reported for the content areas in which they participated. These reporting decisions were made to support the requirement that all students participate in the NECAP testing program.

Data are provided for the following groups of students who may not have completed the entire battery of NECAP tests:

- **Alternate Test**: Students in this category completed an alternate test for the 2005–2006 school year.
- **First-Year LEP**: Students in this category are defined as being new to the United States after October 1, 2005 and were not required to take the NECAP tests in reading and writing. Students in this category were expected to take the mathematics portion of the NECAP.
- **Withdrew After October 1**: Students withdrawing from a school after October 1, 2006 may have taken some sessions of the NECAP tests prior to their withdrawal from the school.
- **Enrolled After October 1**: Students enrolling in a school after October 1, 2006 may not have had adequate time to participate fully in all sessions of NECAP testing.
- **Special Consideration**: Schools received state approval for special consideration for an exemption on all or part of the NECAP tests for any student whose circumstances are not described by the previous categories but for whom the school determined that taking the NECAP tests would not be possible.
- **Other**: Occasionally students will not have completed the NECAP tests for reasons other than those listed above. These "other" categories were considered not state approved.

The results section in the bottom half of the page shows the number and percentage of students performing at each achievement level in each of the three content areas across the school, district, and state. In addition, a mean scaled score is provided for each content area across school, district, and state levels. For the district version of this report, the school information is blank.

The content area results pages provide information on performance in specific subcategories of the tested content areas (for example, geometry, and measurement within mathematics). The purpose of these sections is to help schools to determine the extent to which their curricula are effective in helping students to achieve the particular standards and benchmarks contained in the *Grade Level Expectations*. Information about each content area (reading, mathematics and writing) for school, district, and state includes

- the total number of students enrolled, not tested (state-approved reason), not tested (other reason), and tested;
- the total number and percentage of students at each achievement level (based on the number in the tested column); and
- the mean scaled score.

Information about each content area subcategory for reading, mathematics and writing includes the following:

- The **Total Possible Points** for that category. In order to provide as much information as possible for each category, the total number of points includes both the common items used to calculate scores and additional items in each category used for equating the test from year to year.
- A graphic display of the **Percent of Total Possible Points** for the school, state, and district. In this graphic display, there are symbols representing school, district, and state performance. In addition, there is a line representing the standard error of measurement. This statistic indicates how much a student's score could vary if the student were examined repeatedly with the same test (assuming that no learning were to occur between test administrations).

The disaggregated content area results pages present the relationship between performance and student reporting variables (see list below) in each content area across school, district, and state levels. Each content area page shows the number of students categorized as enrolled, not tested (state-approved reason), not tested (other reason), and tested. The tables also provide the number and percentage of students within each of the four achievement levels and the mean scaled score by each reporting category.

The list of student reporting categories is as follows:

- gender
- Primary Race/Ethnicity
- Limited English Proficiency (LEP)

- IEP
- socioeconomic status (SES)
- migrant
- Title I
- 504 Plan

The data for achievement levels and mean scaled score are based on the number shown in the tested column. The data for the reporting categories were provided by information coded on the students' answer booklets by teachers and/or data linked to the student label. Because performance is being reported by categories that can contain relatively low numbers of students, school personnel are advised, under FERPA guidelines, to treat these pages confidentially.

It should be noted that for NH and VT, no data were reported for the 504 Plan in any of the content areas. In addition, for VT, no data were reported for Title I in any of the content areas.

9.6 SCHOOL AND DISTRICT SUMMARY REPORTS

The *NECAP School Summary Report* and the *NECAP District Summary Report* provide details, broken down by content area, on student performance by grade level tested in the school. The purpose of the summary is to help schools determine the extent to which their students achieve the particular standards and benchmarks contained in the *Grade Level Expectations.*

Information about each content area and grade level for school, district, and state includes

- the total number of students enrolled, not tested (state-approved reason), not tested (other reason), and tested
- the total number and percentage of students at each achievement level (based on the number in the tested column) and
- the mean scaled score.

The data reported, report format, and guidelines for using the reported data are identical for both the school and district reports. The only difference between the reports is that the *NECAP District Summary Report* includes no individual school data. Separate school report and district reports were produced for each grade level tested.

9.7 DECISION RULES

To ensure that reported results for the 2006–07 NECAP are accurate relative to collected data and other pertinent information, a document that delineates analysis and reporting rules was created. These decision rules were observed in the analyses of NECAP test data and in reporting the test results. Moreover, these rules are the main reference for quality assurance checks.

The decision rules document used for reporting results of the October 2006 administration of the NECAP is founded in Appendix L.

The first set of rules pertains to general issues in reporting scores. Each issue is described, and pertinent variables are identified. The actual rules applied are described by the way they impact analyses and aggregations and their specific impact on each of the reports. The general rules are further grouped into issues pertaining to test items, school type, student exclusions, and number of students for aggregations.

The second set of rules pertains to reporting student participation. These rules describe which students were counted and reported for each subgroup in the student participation report.

9.8 QUALITY ASSURANCE

Quality assurance measures are embedded throughout the entire process of analysis and reporting. The data processor, data analyst, and psychometrician assigned to work on the NECAP implement quality control checks of their respective computer programs and intermediate products. Moreover, when data are handed off to different functions within the Research and Analysis division, the sending function verifies that the data are accurate before handoff. Additionally, when a function receives a data set, the first step is to verify the data for accuracy.

Another type of quality assurance measure is parallel processing. Students' scaled scores for each content area are assigned by a psychometrician through a process of equating and scaling. The scaled scores are also computed by a data analyst to verify that scaled scores and corresponding achievement levels are assigned accurately. Respective scaled scores and achievement levels assigned are compared across all students for 100% agreement. Different exclusions assigned to students that determine whether each student receives scaled scores and/or is included in different levels of aggregation are also parallel-processed. Using the decision rules document, two data analysts independently write a computer program that assigns students' exclusions. For each subject and grade combination, the exclusions assigned by each data analyst are compared across all students. Only when 100% agreement is achieved can the rest of data analysis be completed.

The third aspect of quality control involves the procedures implemented by the quality assurance group to check the veracity and accuracy of reported data. Using a sample of schools and districts, the quality assurance group verifies that reported information is correct. The step is conducted in two parts: (1) verify that the computed information was obtained correctly through appropriate application of different decision rules and (2) verify that the correct data points populate each cell in the NECAP reports. The selection of sample schools and districts for this purpose is very specific and can affect the success of the quality control efforts. There are two sets of samples selected that may not be mutually exclusive.

The first set includes those that satisfy the following criteria:

- One-school district
- Two-school district
- Multi-school district

The second set of samples includes districts or schools that have unique reporting situations as indicated by decision rules. This set is necessary to check that each rule is applied correctly. The second set includes the following criteria:

- Private school
- Small school that receives no school report
- Small district that receives no district report
- District that receives a report but all schools are too small to receive a school report
- School with excluded (not tested) students
- School with home-schooled students

The quality assurance group uses a checklist to implement its procedures. After the checklist is completed, sample reports are circulated for psychometric checks and program management review. The appropriate sample reports are then presented to the client for review and sign-off.

SECTION IV
REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). Fort Worth: Holt, Rinehart and Winston.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23,* 355–368.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). New York: John Wiley & Sons, Inc.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105–146). New York: Macmillan Publishing Co.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hambleton, R. K., & van der Linden, W. J. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Joint Committee on Testing Practices (1988). *Code of Fair Testing Practices in Education*. Washington, D.C.: National Council on Measurement in Education.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32,* 179–197.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Muraki, E. & R. D. Bock (2003). PARSCALE 4.1. Lincolnwood, IL: Scientific Software International.
Subkoviak, M.J. (1976). Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement*, 13, 265-276.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589-617.

Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duign, & T. A. B. Snijders (Eds.), *Essays on Item Response Theory,* (pp. 357-375). New York: Springer-Verlag.

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213-249.